Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

Periodic Activity Report

Period covered: from 01/06/2006 to 31/05/2007
Start date of project: June 1st, 2006
Due date of deliverable: May 31st, 2007
Distribution: Public

Date of preparation: 31/05/2007
Duration: 36 months
Actual submission date: May 31st, 2007
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: "Sapienza" Università di Roma
Lead contractor for this deliverable: "Sapienza" Università di Roma

Contents

Publishable executive summary

The vision

TAGora is a project sponsored by the Future and Emerging Technologies program of the European Community (IST-034721) focusing on the semiotic dynamics of online social communities. The widespread diffusion of access to the Internet is making possible new modalities of interaction between Web users and the information available online. The new vision of the Web regards users not only as producers or consumers of information, but also as architects of the information on the Web, which gets shaped according to criteria closely related to the meaning of information, the semantics of human agents. In this perspective the Web is becoming an infrastructure for "social computing", that is, it allows to coordinate the cognitive abilities of human agents in online communities, and steer the collective user activity towards predefined goals.

An approach to information management that has become wildly popular during 2005 (in a matter of a few months), is *collaborative tagging*. The central idea is that users interested in organizing and sharing a certain kind of resources (digital photographs, web pages, academic papers, and so on), use a web application to associate free-form keywords – called "tags" – with the content they're interested in. Such associations are personal, but globally visible to the user community. At the system level the set of tags, though determined with no explicit coordination, evolves in time and leads towards patterns of terminology usage that are shared by the entire user community. Hence one observes the emergence of a loose categorization system – commonly referred to as *folksonomy* – that can be effectively used to navigate through a large and heterogeneous body of resources. Tags act as a sort of "semantic glue" bringing together resources and users in a time-dependent and truly complex architecture, providing an unexpected bottom-up realization of the semantic web vision originally proposed by Tim Berners-Lee.

Overall, the collaborative character underlying many Web 2.0 applications puts them, very naturally, in the spotlight of complex systems science, since the problem of linking the low-level scale of user behavior with the high-level scale of global applicative goals is a typical problem tackled by the science of complexity: understanding how an observed emergent structure arises from the activity and interaction of many globally uncoordinated agents. The large number of users involved, together with the fact that their activity is occurring on the Web, provide for the first time a unique opportunity to monitor the "microscopic" behavior of users and link it to the emergent properties of Web 2.0 applications (for example the global properties of a folksonomy) by using formal tools and conceptual frameworks from Statistical Physics. Understanding how the emergent properties of applications are linked to the behavior of their users is a challenging problem at the interface of several fields, from computer science and complex systems science, to cognitive science and information architecture. TAGora project aims at understanding and modeling information dynamics in online communities, providing a solid scientific foundation for the emerging field of "Web Science".

Scientific and Technological Objectives

The project is articulated in four main areas whose activities are strongly intertwined. The initial phase of the project will deal with collecting actual data from existing, live systems and analyzing

them with a variety of formal tools, eventually inferring models that are able to capture the essential features of the emergent dynamics, and explain how they might arise from the interactions of single agents. The inferred models of the emergent dynamics will be subsequently used to develop simulations that will allow the formulation of design strategies targeted at attaining a specific global behavior.

Emergent metadata The initial phase of the project will deal with collecting actual raw data from existing, live systems. By "raw data" we mean the emergent metadata that arise because of agent interactions in online social communities, as described in the introduction. Several online communities are readily accessible over the web: for a selected set of these systems, tools will be developed and deployed to harvest the relevant data, metadata and temporal dynamics, and to store the acquired information in a form amenable for data analysis.

Data analysis of emergent properties Examining quantitative aspects of folksonomy is a highly important area of research. Our objective is the set up of several protocols of data analysis to be performed on the raw data sets. A data analysis protocol is defined by: (1) indicating a specific quantity / observable / estimator suitable of a quantitative measure on the raw data sets; (2) acquiring the existing software tools, or developing new specific tools, needed to perform the measure; (3) extracting the relevant statistical information characterizing the analyzed data sets.

The aim of the data analysis is to identify and quantify emergent properties of the system in study, i.e. properties that can not be simply inferred from the behavior of the single agent. Beyond suggesting the collection of new or more refined raw data, the results of the data analysis will be used to

- ⋆ identify general features common to the different systems in study
- ⋆ characterize/discriminate the specific features of different systems in study
- ⋆ orient the modelling phase of the research project (see below)
- ⋆ providing benchmarks to test/improve existing systems or to suggest the creation of new more performing systems

Modeling and simulations

The objectives of this research area are twofold:

understanding complexity: develop models that captures the essence of the emergent dynamics and explain how it might arise from the interactions of single agents;

taming complexity: formulate design strategies that allow controlling the behavior of the system at the emergent level by suitably choosing the microscopic dynamics of the interacting agents

One of the most important goals is to construct, implement and study specific modeling schemes aiming at reproducing, predict and control the emergent properties seen in the semiotic dynamics orchestrated in on-line communities. We plan in particular a modeling activity at different scales. On the one hand it will be important to construct microscopic models of communicating agents performing language games without any central control. At a different scale we shall consider more coarse-grained probabilistic models. Several models will be proposed to address specific aspects/scales of folksonomy. The models will allow computer simulation aimed at measuring emergent features to be compared with the results of the data analysis activity. The simulations should give an insight in how users select tags, what kind of categories and category structures underlying the evolving system

of tags, how categories and tags are related to the objects being tagged, etc. It will also give information on what kind of more global structures (such as the most frequent tags) can be provided to users to optimize their on-line community infrastructure. The models will require components for assigning or adopting tags, categorizing data, and collective dynamics. However the approach will be to keep the models as simple as possible, identifying the minimal ingredients responsible for the emergent properties. The minimal character of the models should make a more analytical mathematical study feasible.

A possible way to tackle the complexity of the systems is to individuate different time scales, which can be separated. For instance, we expect that the dynamics of the social network of the folksonomy could be different from the time scale of the dynamics of the resources and of the tags. In this case one can, as a first approximation, propose a model of tags and/or resource dynamics based on a given, slowly evolving social network topology. This kind of assumption should be tested and corroborated as much as possible with the observations coming from the real data analysis.

Feedback and control Finally, the output of all these activities has the potential to feed back into the data collection activity, specifically to the live social tagging system developed as part of, in order to experimentally verify the devised control strategies and demonstrate the technological advantage achieved by the present project.

Long-term applications

Collaborative tagging originated from the need to manage large collections of data. Tagging data is a means to describe, search, and retrieve objects in an intuitive way, which constitutes an important factor of its success. On the long term TAGora will provide experimental systems which are on the one hand intended to further improve navigation possibilities provided by tags, and on the other hand deliver data for the research work of the project. In order to have privileged and controllable data sources for the collaboration, TAGora planned to design and deploy systems - both online systems and actual demonstrations/experiments - for the specific purpose of data collection. The first objective involves building systems that add value to existing tagging sites. One possibility is to enrich navigation based on tags by adding data analysis. The combination of data features and tagging allows to overcome shortcomings of tag-based search, such as problems caused by synonymy, homonymy, missing tags, or spelling mistakes. The added value of original TAGora systems is important in order to attract users and thus fulfill our second objective: to serve as a valuable source for data delivery. Moreover the new applications will allow the Consortium to gain unimpeded access to the raw data and will ultimately provide an experimental "clean room" platform that will be employed to validate the understanding of metadata emergence, and to experiment innovative control strategies. During the first year TAGora deployed prototypical versions of applications. Here are two examples:

BibSonomy BibSonomy (`www.bibsonomy.org`, see Fig. 1), allows users to upload their bookmarks or bibliographic references and assign them arbitrary labels, denoted “tags”. Moreover users may share bookmarks and publication references. In general, social resource sharing systems all use the same kind of lightweight knowledge representation, called folksonomy.

Ikoru Ikoru (`demo.ikoru.net`, see Fig. 2)is a prototypical system developed by Sony CSL that unifies browsing by tags, visual and audio features. This combination allows an intuitive exploration of databases and helps to overcome shortcomings of solely tag-based systems. In contrast to traditional image retrieval approaches, Ikoru employs user tags, complemented by image and music data analysis and classification.

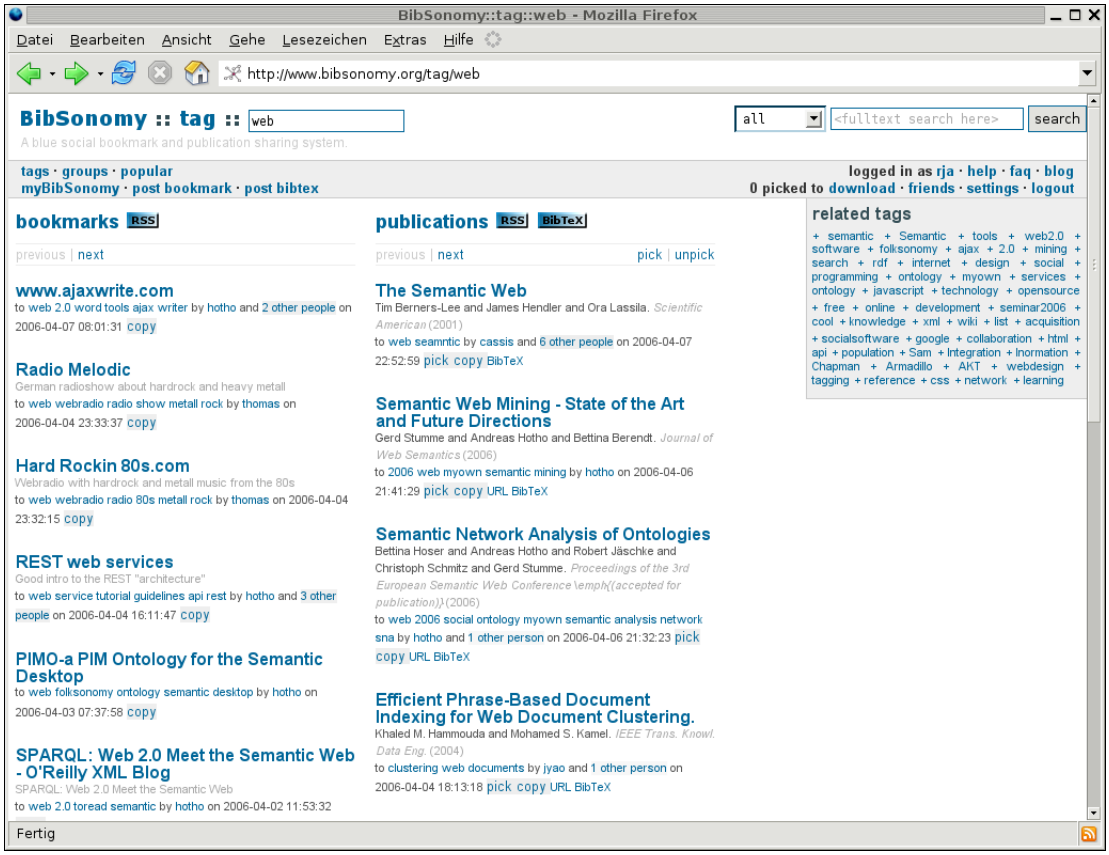


Figure 1: Screenshot of BibSonomy

Results achieved so far

The main results achieved so far include:

i) preparation of a White Paper on open problems and challenges for Semiotic Dynamics in Online Social Communities.

ii) extensive data collection from selected collaborative tagging systems (Full snapshot of del.icio.us and large scale snapshot for Flickr and Last.Fm);

iii) acquisition of existing datasets from several social websites (IMDB, Netflix, Wikipedia);

iv) realization of two web-based applications: BibSonomy (www.bibsonomy.org) and Ikoru (www.ikoru.net);

v) devising new concepts and tools for data analysis;

vi) first stochastic modeling of a collaborating tagging system;

Consortium and contact details

The project is coordinated by Vittorio Loreto (Physics Dept., *Sapienza* Università di Roma) and includes the following partners and node coordinators:

- Physics Department, *Sapienza* Università di Roma (PHYS-SAPIENZA), Italy, Vittorio Loreto

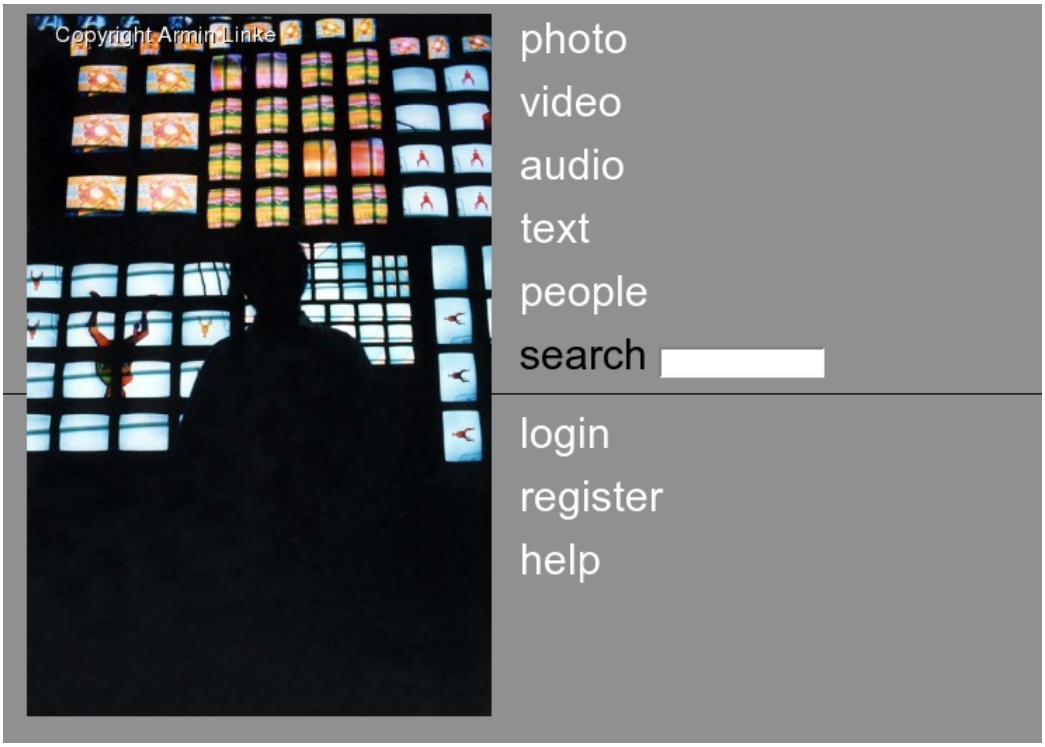


Figure 2: A screenshot of Ikoru from `http://demo.ikoru.net`

- Sony Computer Science Laboratory (SONY-CSL), France, Luc Steels
- University of Koblenz-Landau (UNI KO-LD), Koblenz, Germany, Steffen Staab
- University of KASSEL (UNIK), Kassel, Germany, Gerd Stumme
- University of Southampton (UNI-SOTON), Southampton, UK, Nigel Shadbolt

Please contact:
Vittorio Loreto, Physics Dept., *Sapienza* Università di Roma
Tel: +39 06 4991 3437
E-mail: vittorio.loreto@roma1.infn.it
For more information see: `http://www.tagora-project.eu`

Chapter 1

Project objectives and major achievements during the reporting period

1.1 Project objectives

A new paradigm has gained impact in large-scale information systems: Social Tagging. In applications like Flickr, Connotea, Citeulike, Delicious, etc. people no longer make passive use of online resources: they take on an active role and enrich resources with semantically meaningful information. Such information consists of terminology (or "tags") freely associated by each user to resources and is shared with users of the online community. Despite its intrinsic anarchist nature, the dynamics of this terminology system spontaneously leads to patterns of terminology common to the whole community or to subgroups of it. Surprisingly, this emergent and evolving semiotic system provides a very efficient navigation system through a large, complex and heterogeneous sea of information.

Our project is aimed at giving a scientific foundation to these developments, so contributing to the growth of the new field of Semiotic Dynamics. Semiotic Dynamics studies how semiotic relations can originate, spread, and evolve over time in populations, by combining recent advances in linguistics and cognitive science with methodological and theoretical tools of complex systems and computer science.

The project is exploiting the unique opportunity offered by the availability of enormous amount of data. This goal will be achieved through:

(a) a systematic and rigorous gathering of data that will be made publicly available to the consortium and to the scientific community;

(b) designing and implementing innovative tools and procedures for data analysis and mining;

(c) constructing suitable modeling schemes which will be implemented in extensive numerical simulations.

We aim in this way at providing a virtuous feedback between data collection, analysis, modeling, simulations and (whenever possible) theoretical constructions, with the final goal to understand, predict and control the Semiotic Dynamics of on line social systems.

1.2 Objectives and main achievements of the reporting period

The first main objective of the first year of the project was that to clearly identify challenges and bottlenecks posed by Online Social Communities. The partners had of course already an idea of

what these challenges are but the aim was to make them explicit and discuss them in order to reach a shared view between partners with different and complementary background and expertise and in order to clearly communicate them to the research community at large.

The explicit communication of these challenges among partners with different and complementary background (including physics, computer science, information technology, linguistics) and different expertises (theoretical models, numerical simulations, artificial life and artificial intelligence, ontology learning, semantic web) has proven to be a very effective way to achieve this inter-disciplinary discussion. This discussion, in turn, has lead to the preparation of the White Paper on *Target Problems and Grand Challenges for the Semiotic Dynamics in Online Social Communities (deliverable D5.3)* that involved all partners and that, we believe, could become a reference point in this research area.

The second main objective of the year was to start making progresses along the main research directions of the project, namely:

- **infrastructures** developing tools and deploying data collection infrastructures (software, servers, network connectivity) for gathering data from collaborative tagging systems;
- **data analysis** devising methods and algorithms for analyzing the raw data from the data-collection campaign;
- **modeling** developing theoretical constructions and models whose outcomes have to be compared with the experimental findings;
- **applications** developing and making publicly available innovative applications embodying novel navigation and control concepts;

Details of progresses achieved in these research lines are provided in Chapter 2. Here we only emphasize that significant progresses have been made in the three following areas:

A **data collection** All the partners participated to an extensive data collection campaign from selected collaborative tagging systems. In particular the Consortium crawled `del.icio.us` and Flickr. Because of the large size of the chosen systems, crawling has to be done in a distributed fashion, with several machines crawling the target systems and storing data on a central server. Finally, the gathered data have to be post-processed in order to assess data-quality, remove inconsistencies and encode them in a way which is more amenable to further analysis.

B **modelling activity** The PHYS-SAPIENZA team succeeded in proposing the first stochastic model to explain users' activity on collaborative tagging systems. Our stochastic model is meant to describe the behavior of an "effective" average user in the context identified by a specific tag, and can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of a small number of words. Fitting the parameters of the model, in order to match its predictions (obtained by computer simulation) against the experimental data, we obtain an excellent agreement for all the frequency-rank curves measured. This is a clear indication that the tagging behavior embodied in our simple model captures some key features of the tagging activity. The above results were have been published in the Proceeding of the National Academy of Sciences (Cattuto et al., 2007) and featured in the news, both in scientific and non-technical journals and magazines.

C **applications** Two main online applications have been deployed by the Consortium. *BibSonomy*, which allows users to manage their bookmarks or bibliographic references by assigning

them arbitrary labels, denoted “tags”. The dataset includes now data from approximately 400 users, 12.000(different)/140.000(all) tags and 39.000 resources and can easily be loaded into a mysql data base. Very recently a special BibSonomy webpage has been set up (`http://www.bibsonomy.org/events/statphys23/`) as a friendly interface for all the participants to the XXIII International Conference on Statistical Physics (STATPHYS 23) to be held in Genova, Italy on 9-13 of July 2007. All contributions submitted to the Statphys23 conference have been loaded into the system, together with the keywords (tags) that authors have associated with their abstracts. In this way BibSonomy will serve a main navigation tool through the sea of abstracts of STATPHYS 23. *Ikoru* is a prototypical system developed by Sony CSL that unifies browsing by tags, visual and audio features. This combination allows an intuitive exploration of databases and helps to overcome shortcomings of solely tag-based systems. In contrast to traditional image retrieval approaches, Ikoru employs user tags, complemented by image and music data analysis and classification.

Chapter 2

Workpackage progress of the period

Following are the objectives that were planned for all the workpackages for the first year of activity of the project.

2.1 Workpackage 1 (WP1) - Emergent Metadata

2.1.1 Objectives

The goal of WP1 is to collect raw data about the static and dynamical properties of folksonomies and, based on such data, to identify stylized facts (emergent features) for subsequent investigation and modeling. The activity of data collection covers the design and development of software clients, the deployment of clients, and the post-processing of raw data to produce high-quality large-scale datasets to be used for data mining.

Task 1.1 Data from collaborative tagging (folksonomies)

The goal of this task is to provide the project with large-scale and well documented datasets from existing folksonomies. These data are meant to be a foundation for an extensive scientific investigation of the statistical properties of folksonomies, as well as to be used to gain insight into user behavior and tagging patters. The objectives of this task comprise the identification of suitable data sources, the development and deployment of software clients for data collection, the post-processing of data, and the assembly of data repositories for the project, and related documentation.

The work of Task 1.1 initially focused on two extremely popular system: *del.icio.us* (`http://del.icio.us`), a social bookmarking system, and *Flickr* (`http://flickr.com`), a photo-sharing system. Both systems were among the very first to deploy and diffuse the ideas of collaborative tagging, they have a very large user base, and as a consequence, they are regarded as paradigmatic folksonomies.

The information that is relevant for data analysis broadly consists of the following categories:

- the list of tagged resources (bookmarked web pages, digital photographs), as well as unique identifiers that allow to track them within the target system
- the set of tags associated with a given resource
- the set of users, and the social network of their relations within the system, when applicable
- the set of resources associated with a given tag

- the temporal series describing the evolution of the system in terms of inserted resources, added tags, new users and so on.

All the above information is available through the web interface of the corresponding system, but no access is given to the tagging data in a structured form, so a crawling strategy has to be devised. Because of the large size of the chosen systems, crawling has to be done in a distributed fashion, with several machines crawling the target systems and storing data on a central server. Finally, the gathered data have to be post-processed in order to assess data-quality, remove inconsistencies and encode them in a way which is more amenable to further analysis.

Task 1.2 Data collection from the bibliographic reference sharing system of the Consortium BibSonomy

In the beginning of 2006 the Institute of Knowledge Engineering and Discovery introduced a novel service to manage and share web pages and publications in the Internet at www.bibsonomy.org. User of the open access system can centrally store bookmarks for web pages in form of URLs and add keywords to those resources. The keywords, so-called tags, permit users to structure their collection of bookmarks as well as to find already posted entries. The structure resulting from making public each users' individual classification is called a 'folksonomy'.

BibSonomy offers users to search bookmarks of other participants with similar interests and to discover new, interesting web sites. This social perspective allows for personalized recommendations; a feature global search engines such as Google can not provide as they are not aware of their searcher's interests. Due to the central storage of bookmarks users can access their bookmarks from any computer any time.

A primary challenge of scientific work is the search of literature. However, support for a systematic structuring of retrieved publications is rare: typically researchers develop their own management and storage system, and each research group commits high efforts to manage and update publication lists. When the chair of Knowledge Engineering and Discovery was discussing a solution for its professional literature management, the group decided to add to its social bookmark sharing system the possibility of collective literature management as a second, essential component. The institute offers interested communities (e.g. university groups or groups of the German society for computer science) the creation of user groups in BibSonomy to enable further teams to organize their internal and external literature exchange.

Apart from the established publication metadata such as author, title, publisher, year etc. freely chosen tags can be added to the description of documents. Librarians proceed in a similar way when they gather new books and provide descriptive metadata for refinding them. While the librarian keywords are part of a pre-specified, firm vocabulary, users in BibSonomy are free to select their tags. A non-hierarchical, spontaneous handling of vocabulary creation can describe (anticipate?) actual themes which a traditional catalogue (yet?) can not capture. In contrast, the structured librarian organization of keywords does not face difficulties when using terms with multiple meanings. However, librarians are trained to add systematic metadata which can not be assumed for users of freely available literature organization systems. A folksonomy offers the possibility to add more than one tag to a resource, so that documents can be found following different search paths; a book in a library is placed in one physical location. The bigger flexibility of tagging in bookmarking systems is redeemed by the difficulty of finding information. The institute has developed an algorithm to improve the search in folksonomies.

The BibTex format was selected to store the publication metadata. BibTex allows for an integration of reference entries into the document preparation system LaTex with which researchers, especially in natural science, format written work. BibSonomy also creates literature lists in further formats to free researchers from unnecessary work when using different text publication systems. For instance, an institute's publications list often needs to be retyped in different formats to meet the

formatting requirements of the institute's web site, the annual report or other reports for research projects. Using BibSonomy, the literature list is once centrally stored and can then be exported in the required format. Besides BibTex, EndNote often used for publications written in MS Word, as well as formats such as XML, RSS-Feeds and formated HTML are supported.

BibSonomy is one of several bookmarking systems created in the advent of the web 2.0 - the next generation of the world wide web. Other bookmarking systems allow for the administration of photos and music - even the exchange of one's personal goals is possible. Up to date, BibSonomy is the only system combining the management of bookmarks and publications. It was developed to be used by scientists; feedback of many research colleagues has contributed to its design. On 20 February, 2007, the system comprised of 3295 registered users, 149701 bookmarks and 27411 publications. Additionally, 841245 publications have been automatically transferred from the computer science library of the University of Trier. The system has about 250000 page views per day; the tendency is increasing exponentially.

An essential requisite of implementation was the realization of an efficient response to system queries. More details about the implementation and the system's functionality are described in the attached publication (Hotho et al., 2006a). At this place, we welcome you to explore the system at `http://www.bibsonomy.org`.

BibSonomy existed in a rudimentary version already at the beginning of the project. Within TAGora, we have improved its functionality (as discussed in Deliverable 2.1) for attracting a larger user base. The folksonomy data of BibSonomy are available to the consortium for modeling and analyzing interactions in online social communities, and also to other researchers for evaluating folksonomy-centered knowledge discovery and information retrieval algorithms.

Task 1.3 Data from experimental tag-based navigation system at SONY-CSL

The tagging system for images and music developed by Sony CSL (see Task 2.2) will be put online for use by the general public. The system will also serve as backend in user studies, artistic projects, and small-scale projects with selected communities. We aim to gather data from these deployments and use it as a source for the analysis tasks to be carried out by members of the TAGora project. Because we can store detailed information about the web site's visitors (for example, their browsing history), the partners can potentially test hypotheses in greater detail than with the limited data obtained from proprietary tagging systems.

Task 1.4 Collecting data from online recommendation systems

Many e-commerce systems that exist on the web provide some sort of recommendations. One example is Last.fm[1], an online radio station and community web site for listening to and managing music profiles. Users can create music profiles and then automatically record the music they listen to (using the Audioscrobbler plugins) over time to build a model of their listening habits. Users can also tag songs, artists and albums to denote their musical interests and share playlists with their friends. By comparing listener's profiles to find overlaps in taste, last.fm is able to recommend new artists and songs to users based on the listening habits of their *musical neighbours* - those that have similar tastes. With large amounts of data available on the songs users have listened to, last.fm are able to generate unique music charts showing the most popular bands and songs in terms of their global play-time. Official music charts, which are based on sales, can be retrieved from other sources, such as top40-charts[2], and used as a basis for comparison with the last.fm charts. Such a study can be used to explore the evolution of tags and how these tags might

[1] `http://last.fm`
[2] `http://www.top40-charts.com`

influence, or be influenced by, the general music market. It will also be used to investigate semantic recommendation techniques.

Netflix[3] is an example of another e-commerce recommender system. Netflix is an online DVD rental service, established in 1998, that provides a flat rate, mail-based, rental service to customers in the United States. Their current DVD collection contains around 75,000 titles, offered to a customer base of over 6 million individuals. After renting a movie, customers may enter their rating of the movie into the Netflix database via the website, using a discrete score from 1 to 5. In October 2006, Netflix began a competition to find better recommendation systems, offering a grand prize of $1 million to anyone managing to improve on their own algorithm by 10%. To drive this competition, Netflix published a large set of movie rating data from their database featuring 480,189 customers and 100,480,507 ratings across 17,770 movie titles. By combining Netflix data on rentals and ratings with movie information and tagging data from the Internet Movie Database[4], we can explore better recommendation strategies with a solid data set (from Netflix) to evaluate our methods.

2.1.2 Progress

Task 1.1 Data delivery from selected folksonomy

Del.icio.us benchmark dataset Within the project, we crawled also the prominent social bookmarking system del.icio.us. These are available only for use within the TAGora project.

The del.icio.us data have been crawled from November 10 till 24, 2006. The crawling was supported by all participants of the TAGora project and coordinated by the University of Kassel. The crawl was coordinated by a central server in Kassel. It monitored the 'recent posts page', resulting in an constantly updated list of user names. From this list, it distributed small chunks to over nearly 70 PCs, located all over the world, especially at Kassel, Rome, Koblenz, Southampton, Japan, Netherlands, Karlsruhe and Leipzig. These PCs crawled completely the corresponding user pages (including all follow up pages when a user page surpassed 5000 entries). Each PC waited a random time between 60 and 120 seconds before the next request, with an average delay of 90 sec. Globally, this resulted in an average of one request per second on the del.icio.us server. User pages that could not be downloaded completely were marked as 'incomplete' in the database.

Interested project members can contact Andreas Hotho (hotho@cs.uni-kassel.de) to get access to the dataset in html format via the server on `http://www.kde.cs.uni-kassel.de/crawldataset/`. The crawled data have also been made available in HDF5 format by the University of Rome and are accessible on sismopil.phys.uniroma1.it. Interested people can contact Ciro Cattuto (ciro.cattuto@roma1.infn.it) to get access to the dataset.

Overall, the project members have internal access to 10GB compressed and 50GB uncompressed data to analyze and model the evolution and the behavior of social resource sharing systems. Overall, data from 667,128 users of the del.icio.us community with 18,782,132 resources, 2,454,546 tags (organized in 667 bundles) and 140,333,714 tag assignments were collected.

Flickr benchmark dataset A crawler for the Flickr data was developed in Koblenz. The approach of the Flickr crawler is similar to the del.icio.us crawler but tailored to the peculiarities of Flickr. It is implemented in Java and uses an open source java library to directly access the Flickr API. The backend for the crawler is a Postgress database that stores the essential tagging information including the user-tag-photo relations and some additional information returned by the Flickr API.

[3] `http://www.netflix.com`
[4] `http://www.imdb.com`

Java was chosen as implementation language for the Flickr crawler because it can be run on any platform and allows for easy access to databases and the Flickr API via freely available libraries. The crawler strategy allows to run the crawl in parallel on distributed machines for predefined time intervals. The implementation is extensible to allow for easy adaption to new requirements for future crawling activities.

Extensive test with the Flickr API were conducted beforehand to ensure that all required data is correctly retrieved and a consistent dataset will be created. Some problems with the Flickr API were detected and workarounds implemented to avoid possible data irregularities. The actual crawl was then supervised by Koblenz and supported with the necessary infrastructure.

Helpful discussions with partners, especially with Kassel, concerning the crawling strategy and crawled data helped to improve the Flickr crawler.

Due to the huge amount of data obtained from Flickr the crawl has been initially limited the time frame 2004-2005. Subsequent crawls shall extend the dataset - ideally retrieving all tagging data until the current date. Overall, until 15th May 2007 data from 298,954 users of the Flickr community with 24,599,875 resources, 1,553,253 tags and 110,345,103 tag assignments were collected.

Task 1.2 Data collection from the bibliographic reference sharing BibSonomy

BibSonomy benchmark dataset BibSonomy has been announced on several mailing lists to attract more users. The list of mailing lists includes:

- dbworld
- kdnet-members@iais.fraunhofer.de
- wi@aifb.uni-karlsruhe.de
- ak-kd-list@aifb.uni-karlsruhe.de
- fgml@cs.uni-kassel.de
- fg-db@informatik.uni-rostock.de
- fca-list@aifb.uni-karlsruhe.de
- orgmem@aifb.uni-karlsruhe.de
- dl@dl.kr.org
- kaw@science.uva.nl
- community@mlnet.org
- web_graph_algs@yahoogroups.com
- webir@yahoogroups.com
- ontoweb-list@lists.deri.org
- semanticweb@yahoogroups.com
- seweb-list@lists.deri.org
- cg@conceptualgraphs.org
- kweb-all@lists.deri.org

- all-prolearn@agws.dit.upm.es
- ml@isle.org
- AI-SGES@JISCMAIL.AC.UK
- machine-learning@yahoogroups.com
- mlearn@googlegroups.com
- Web-Mining@googlegroups.com
- Machine-Learning@googlegroups.com
- Data-Mining@googlegroups.com
- INDUCTIVE@LISTSERV.UNB.CA

The anonymized data of BibSonomy are downloadable via a mysql dump, which will be updated every half year. It includes the folksonomy, i. e., a list of triples of the form (user, tag, resource), where the resources are either bookmarks (URLs) or references to publications (in BibTEX). Interested people get an account from Miranda Grahl (mgr@cs.uni-kassel.de) for access to our server on `https://www.kde.cs.uni-kassel.de/bibsonomy/dumps/2006-12-31.tar.gz`. Before starting the download, participants have to sign a license agreement in which terms of use are set up. The dataset includes data from approximately 400 users, 12.000(different)/140.000(all) tags and 39.000 resources and can easily be loaded into a mysql data base.

To attract more users, we are currently adding all approx. 1000 abstracts of the XXIII IUPAP International Conference on Statistical Physics [5] to BibSonomy, with the aim to extract physicists as new community. The collection of abstracts will be made available as separate dataset.

Task 1.3 Data from experimental tag-based navigation system at SONY-CSL

A first dataset was gathered during an initial user study involving students of IUAV (Universitá di Venezia) engaged in the task of tagging photographer Armin Linke's images. Thirty students were asked to associate tags to 4600 photos made by Armin Linke. This process resulted in the creation of 300 different tags, and 4102 tag assignments. A version of Ikoru installed in a server at the University was used for this experiment, accessible at `http://armin.iuav.ikoru.net`. The collected data can be made available to other members of Tagora upon request, in the format of SQL and XML files, or directly through the Ikoru API. The dataset of the collective image-tagging experiment was the basis for seven books that were displayed at the Intensive Science exhibition (October 6-7, Paris), organized by SONY-CSL. Each book contained a set of pictures corresponding to one of the tags that the students used in the experiment (see also Milestone 5.2).

The collaboration with Armin Linke continued after this first experience. We are currently finalizing an installation that will go on display in the Zentrum fur Kunst and Medientechnologien (ZKM) in Karlsruhe, Germany. This installation aims to be a physical implementation of a tagging site. The visitors of the exhibition will be able to browse thousands of paper copies of Armin's photos and pick out a selection. The chosen photos can than be assembled into a book that will be printed out at the exhibition using an editing table that was specifically designed for this exhibition. Before the book is printed, the visitors are asked to enter a single word as a title of their collection. This title together with the selected photos are stored in the Ikoru database as tag assignments. A visualization of the state of the database and the evolving relations between the tags will be shown as part of the exhibition. Parallel to this physical installation, a web site will designed that extends the activity of the exhibition to the Internet.

[5] `http://www.statphys23.org/`

The Ikoru platform serves as a backend for both the installation and the web site. The data collected through this experiment will be accessible to the other partners. The exhibition is due to start in October 2007. Two videos of this work in progress is available at `http://www.csl.sony.fr/~hanappe/ArminLinke-HfG-Karlsruhe/`.

Task 1.4 Collecting data from online recommendation systems

The University of Southampton have established contact with one of the directors of Last.fm and organized two informal meetings with them to encourage them to provide data and gain their interest. Last.fm, the company, is interested in the research being carried out in TAGora, and has already provided us with some useful data containing Last.fm user account IDs and the URL for their Flickr or del.icio.us accounts. More Last.fm data has been crawled by TAGora, and currently consist of all the top 50 bands that each of a collection of 7000 Last.fm users listened to, as well as information about 73 albums, 200 tags, 1500 artists, 65000 tracks, and 18000 sound extracts (26 seconds each). More data will be collected from this folksonomy over the few coming months.

Top40-charts is a web site that provides weekly results of music charts for various countries. We have developed scripts to harvest the chart results for UK albums on weekly bases, and store the output in RDF triples in a triple store (3Store (Harris and Gibbins, 2003)). This data will be used along with Last.fm to try and detect any correlations between the two communities. Currently we hold data about the UK top 20 album charts for every week since January 2005. The data is stored in a triple store and accessible with SPARQL queries.

Data from IMDB is made available for download from the main web site in a flat-file, text-based format. We parsed the data and created a relational database of movies, actors, and production crew. This database currently holds nearly 900,000 movies, 2,564,990 names (actors, writers, directors, etc), and 32,247 keywords (1,166,149 keyword to movie assignments).

By mapping this relational database to an ontology of movies, we were able to utilize the D2RQ (Seaborne and Bizer, 2004) mapping technology and provide a SPARQL end-point that supports semantic querying. This approach helps to overcome the scalability limitations of existing triple stores, while maintaining high performance.

Netflix has recently released a large amount of their data for research on recommendation techniques, has also been collected, stored in a relational database, and mapped to an ontology using D2RQ. We stored 480,189 customers, 100,480,507 ratings, and 17,770 movies, Figure 2.1 shows the ontology of movies that we use to provide a homogeneous view over both data sources.

We have already started experimenting with different recommendation techniques that make use of this integration to provide better movie recommendations to Netflix users. The Netflix data is historical data. In our experiments, we generate recommendations (can be thought of as predictions in this case) of which movies the users will like (rate high) or dislike (rate low) based on part of the data (say 2004-2005), then evaluate our results against the actual ratings that took place after this period (Szomszor et al., 2007).

2.1.3 Milestones

M1.1 (Task 1.1) Implementation of software clients and hardware infrastructure to perform data collection from folksonomy sites (month 5).

Small-scale Del.icio.us dataset The work of data analysis and modeling of the PHYS-SAPIENZA team has started well before the availability of the large-scale dataset from the distributed crawl led by the UNIK team. During the very early phase of the projects, a Python-based software client to extract data from Del.icio.us was designed at developed. The client allows to focus on a given tag, resource, or user and download all the related posts. Given a resource or

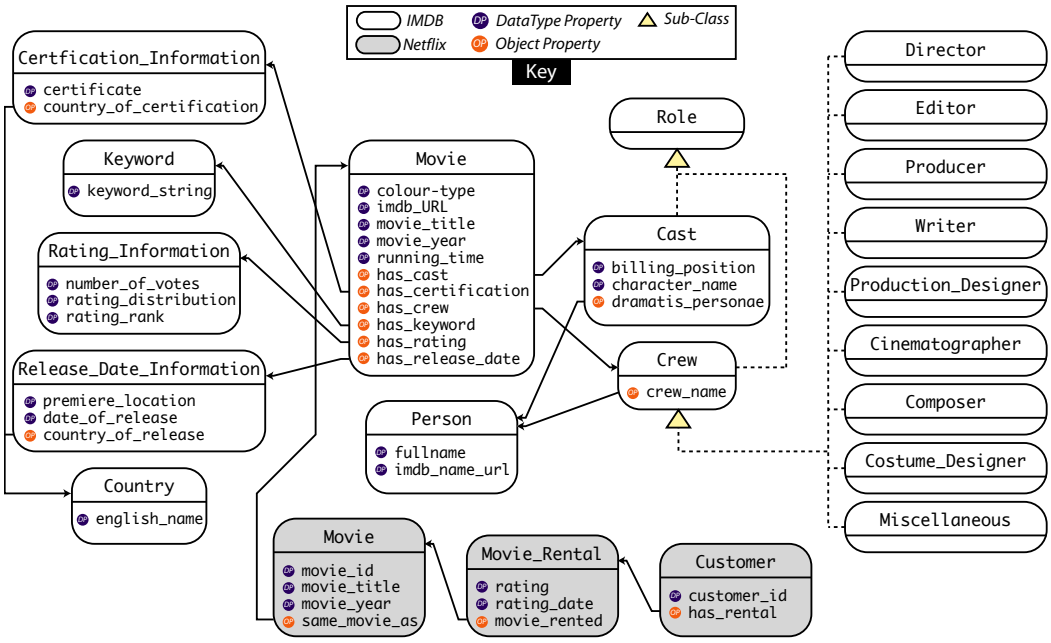


Figure 2.1: The ontology used to integrate IMDB and Netflix data.

a user, all posts involving it can be recovered, while given a tag, only the most recent 10^4 posts can be usually recovered because of a limitation imposed on the Del.icio.us web interface. Small scale datasets for exploratory analysis were built on a per-project basis and stored as serialized Python objects for easy post-processing. The use of this client has been discontinued as soon as the large-scale Del.icio.us datasets became available, but is still available for lightweight forms of data collection.

BibSonomy benchmark dataset A SQL query has been set up to generate the half-yearly dumps of the BibSonomy folksonomy. The most recent dump can be found at `https://www.kde.cs.uni-kassel.de/bibsonomy/dumps/2006-12-31.tar.gz` (see details above), the next dump is scheduled for end of June 2007.

Del.icio.us benchmark dataset As described above (Progress Task 1.4), the del.icio.us data have been crawled from November 10 till 24, 2006. The algorithms have thus been developed – and even deployed – before the milestone.

Flickr benchmark dataset A platform-independent Flickr crawler has been developed in Java on top of a library providing direct access to the Flickr API. A suitable crawling strategy was implemented that retrieved all photos for all known tags in predefined time intervals. A Postgres SQL database is used as backend to store the huge amount of crawled data. The crawling started at the beginning of March. Although all partner dedicate machines for this distributed crawl it is still ongoing due to the large amount of tagging data available from flickr. Therefore the currently crawled dataset will also be limited to the period until the end of 2005. Subsequent crawls will continue to crawl the consecutive years. In case of new requirements the crawler can be easily adapted to adapt to the changed crawling strategies. The final dataset will be directly made available for all partners. Until 15th May 2007 data from 298,954 users of the Flickr community with 24,599,875 resources, 1,553,253 tags and 110,345,103 tag assignments were collected.

SONY-CSL The Last.fm dataset obtained by Sony CSL was obtained through parsing the site's HTML pages (screen scraping). For this, a collection of standard Unix tools were used: *bash*, *wget*, *grep*, *awk*, and *xsltproc*. For the Flickr.com data, a small Java application was written that uses the Flickr API. This application pulls in the information of the tags and photos of a given user account and converts it to SQL queries. The images are downloaded and stored on the local disk (only the medium-sized photo is downloaded).

M1.2 (Task 1.4) Design and deploy a centralized system for storing the selected online resources (month 5).

A centralized system for the storage of all the data used by the TAGora project is impractical due to the extensive scale and size of the data, and the current frequency of update of these data collections that are being collected by a number of partners in the consortium. Also, we would like to have the ability to experiment with distributed knowledge-base systems, which requires some of the data stores to remain separate. Note that in some cases, when heavy analysis is required, it might be best to have a local copy of the data to ensure maximum access speed.

However, part of our research will investigate methods and techniques to integrate some folksonomy data, and hence it is important for all our data stores to provide external access to local data repositories. For our data collections that will be used for work on recommendation systems, such as Netflix and IMDB, we store the data in a centralized server. We designed and used ontologies to represent the data collections, and mapped together some of the classes and individuals in the knowledge bases using `owl:sameAs` properties, thus semantically bridging the data.

2.1.4 Deviations and Corrective Actions

For milestone M1.2, the Consortium we decided not to create a centralized data repository for the whole consortium due to the current frequent changes made by the TAGora partners to the data collections. However, Southampton currently hosts a powerful server where all collected data is stored in triple stores or relational databases, and accessed via SPARQL queries or web browsing interfaces. More data sets will be added to this server when required.

UNI-SOTON At the beginning of the project, UNI-SOTON faced difficulties recruiting a good quality candidate which delayed the task of data gathering by a few months. As a corrective action, Harith Alani, one of the local TAGora co-PIs dedicated 50% of his time for this project and built some of the needed tools and repositories to gather and store some of the data we require. UNI-SOTON plans to hire one or two final year students over the summer of 2007 to help gathering more data from specific resources (e.g. Last.fm).

2.1.5 Deliverable and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
1.1	Data delivery from selected folksonomy sites.	1	31 May 2007	31 May 2007	10	10.99	**PHYS-SAPIENZA**
1.3	Data delivery from selected recommender systems.	1	31 May 2007	31 May 2007	6	5	**UNI-SOTON**

Mil. No.	Milestone name	WP No.	Date due	Actual/ Forecast delivery date	Lead contractor
M1.1	(Task 1.1) Implementation of software clients and hardware infrastructure to perform data collection from folksonomy sites.	1	31 October 2006	31 October 2006	**PHYS-SAPIENZA SONY-CSL UNI KO-LD UNIK**
M1.2	(Task 1.4) Design and deploy a centralised system for storing the selected online resources.	1	31 October 2006	31 October 2006	**UNI-SOTON**

2.2 Workpackage 2 (WP2) - Applications

2.2.1 Objectives:

Task 2.1.1: Folksonomy website for sharing of bibliographic data.

The objectives of the work carried out during the first year of the project in this WP is to obtain first prototypes of the different systems, so that data collection can start in Year 2. The progress in the development of the different systems is discussed briefly below. For more details, we refer to the deliverables D 2.1 and D 2.2.

For BibSonomy, a preliminary version of this social tagging system for bibliographic data was already online at the beginning of the project. The objective in the first year of the project was to extend its functionality in order to increase its user base, and to provide easy access to the data.

Task 2.1.2: Folksonomy peer-to-peer system for sharing of bibliographic data.

The main objective of this task is to build a distributed collaborative tagging systems which is basically a decentralized version of common collaborative tagging services like flickr, delicious etc. Therefore it will generally provide the same functionality and features like those systems which is the ability to tag personal information objects (i.e. photos, documents, videos, etc.) and publically share the tagging metadata. Additionally, tagging statistics will be made available like the commonly used tag clouds which show the different use frequencies of tags.

The peer-to-peer tagging system is inspired by Bibster (Haase et al., 2004), an award-winning peer-to-peer system for sharing bibliographic data employing a generic semantic peer-to-peer platform based on Sun's JXTA. But due to the limitations of a centralized ontology used for annotation purposes it does not allow arbitrary tags as content categories. A new architecture (SEA) (Franz et al., 2006) with more flexible content tagging capabilities overcomes those limitations and provides means for easy sharing of tagged resources on top of a peer-to-peer technology.

The main focus of the peer-to-peer tagging system will be on multimedia data which usually already contains rich metadata itself (e.g. ID3 or EXIF metadata in mp3 files or photos respectively). The system will therefore provide means to automatically extract such metadata and to easily organize and manipulate the stored information through a simple to use editor interface. This will also include the ability to easily navigate via tags across the media boundaries.

An implementation of the peer-to-peer tagging system will be run within a case study. The collected test data will be structurally similar to those gathered from flickr, but tagging behaviors may differ because the peer-to-peer system works completely decentralized.

Task 2.2 Tag-based navigation systems

Sony CSL developed a prototypical tagging system for music and images. The goals of this work were multiple. First, we wanted to explore how the tag-based navigation could be improved with the use of data analysis. Second, we want to be able to register the detailed browsing history of the visitors in order to analyze possible relations with the tagging data. Third, we integrated this work in a reusable software component using Web standards so that the research effort can be available to other interested parties. Lastly, the tagging system is deployed as a Web site, but is also used in specific, artistic projects. This has allowed us to gather data that will be shared with the other partners.

2.2.2 Progress

Task 2.1 Social tagging for online scientific communities

Task 2.1.1: Folksonomy website for sharing of bibliographic data

This section briefly describes the BibSonomy system[6] developed by our group. After an introduction to the architecture of BibSonomy, we explain features that we have added within the project.

Architecture The basic building blocks of BibSonomy are an Apache Tomcat[7] servlet container using Java Server Pages[8] and Java Servlet[9] technology and a MySQL[10] database as backend.

Currently the project has several thousand lines of code and is using the Model View Controller (MVC) (Krasner and Pope, 1988) programming paradigm to separate the logical handling of data from the presentation of the data. This enables us to produce output in various formats (see Section 2.2.2), since adding a new output format is accomplished by implementing a JSP as a view of the model.

The database schema of BibSonomy is based on four tables: one for bookmark posts, one for publication posts, one for tag assignments (*tas*) and one for *relations*. Two further tables store information regarding *users* and *groups*. In Figure 2.2, the two *posts* tables are shown as one and it is only hinted that these are really two tables. The reason to show them as one table *posts* is that they're very similar – the publication post table has just some additional columns to hold all the BibTEX fields. They are separated in the database for efficiency reasons, since these extra columns just need to be stored for publications.

The posts table is connected with the tas table by the key *post_id*. The scheme is not normalized, on the contrary we have added a high amount of redundancy to speed up queries. For example, besides storing group, user name and date in the posts table, we also store it in the tas table to minimize the rows touched when selecting rows for the various views. Furthermore several other tables hold counters (i.e. how many people share one resource, how often a tag is used, ...). Finally a lot of indexes (12 in the tas table alone) build the basis for fast answering of queries.

Overall we spent a lot of time investigating and optimizing SQL queries and table schemes and tested both with folksonomy data of up to 8.000.000 posts. At the moment we need no special caching or physical distribution of the database to get reasonable response times, although the system is scalable, since distribution of queries over synchronized databases is possible with MySQL.

[6] http://www.bibsonomy.org
[7] http://tomcat.apache.org
[8] http://java.sun.com/products/jsp
[9] http://java.sun.com/products/servlets
[10] http://www.mysql.com

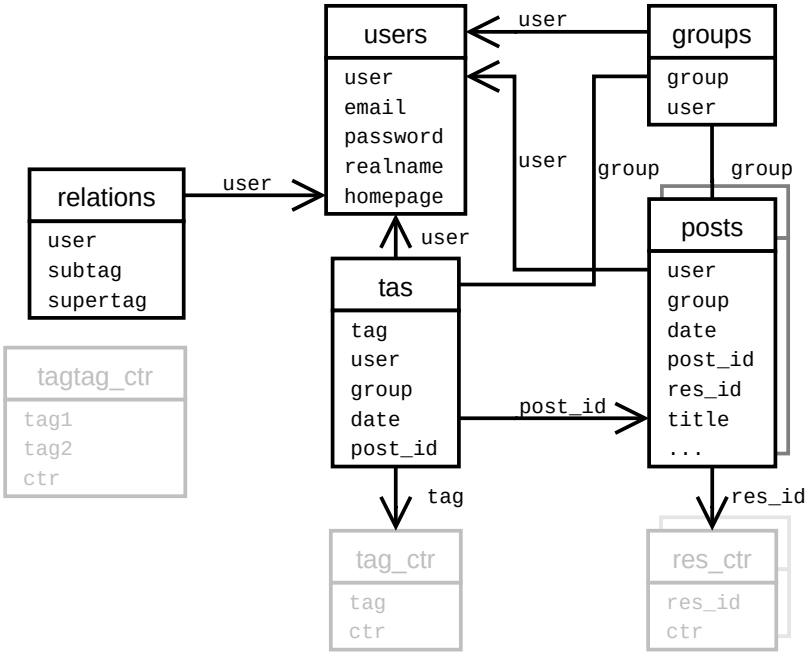


Figure 2.2: Relational schema of the most important tables.

Features This section describes some extensions of BibSonomy which were not part of the basic system but turned out to be necessary for the everyday use of BibSonomy.

Relations between tags Tagging gained so much popularity in the past two years because it is simple and no specific skills are needed for it. Nevertheless the longer people use systems like BibSonomy, the more often they ask for options to structure their tags. A user specific binary relation $\prec$ between tags is an easy way to arrange tags. Therefore we included this possibility in BibSonomy.

To enable the addition of elements to the relation already during tagging, we decided to reserve the character sequences <- and ->. That means, if the user u enters t_1->t_2, we attach the tags t_1 and t_2 to the respective resource and add the triple (u, t_1, t_2) to the relation $\prec$. The tag t_2<-t_1 is interpreted as t_1->t_2. Consequently it is not possible to have tags which contain the strings "<-" or "->". The semantics of this relation can be read as "t_1 *is a* t_2" or "t_1 is a *subtag* of the *supertag* t_2". There are also other ways to add elements to $\prec$, in particular a relation editor.

Usage of this relation is made in several situations. First, the user can structure his tag cloud by showing all subtags of a certain supertag and therefore can see the tags in a hierarchy. Second, BibSonomy offers the option to show on a users tag page not only posts which contain a certain tag, but also posts which contain one of the subtags of the specific tag. This works also for tag intersections. Compared to *tag bundles* which are available in del.icio.us, this relation is more general and more powerful.

Bringing this relation into the system raises several questions which are still under discussion:

- How to handle cycles, i.e. $u \in U$ and $t_1, \ldots, t_m \in T$ with $(u, t_i, t_{i+1}) \in \prec$ (for $i = 1, \ldots, m-1$) and $(u, t_m, t_1) \in \prec$?
- How to model equivalence or non-equivalence of tags?
- Should we make use of the transitive closure of the relation? If so: where and how to do it efficiently?

- How to express such queries like “all posts which have the tag *howto* and also one of the subtags of *programming*"? One idea involving the reserved character sequences is “`->programming howto`".

Duplicate detection In particular for literature references there is the problem of detecting duplicate entries, because there are big variations in how users enter fields such as journal name or authors. On the one hand it is desirable to allow a user to have several entries which differ only slightly. On the other hand one might want to find other users entries which refer to the same paper or book even if they are not completely identical.

To fulfill both goals we implemented two hashes to compare publication entries. One is for comparing the entries of a single user (*intra* user hash) and one for comparing the entries of different users (*inter* user hash). Comparison is accomplished by normalizing and concatenating BibTEX fields, hashing the result with the MD5 (Rivest, 1992) message digest algorithm and comparing the resulting hashes. MD5 hashing is done for efficiency reasons only, since this allows for a fixed length storage in the database. Storing the hashes along with the resources in the posts table enables fast comparison and search of entries.

The intra user hash is relatively strict and takes into account the fields *title*, *author*, *editor*, *year*, *entrytype*, *journal* and *booktitle*. This allows one to have articles with the same title from the same authors in the same year but in different volumes (e.g. a technical report and the corresponding journal article).

In contrast, the inter user hash is less specific and includes only the *title*, *year* and *author* or *editor* (depending on what the user has entered).

In both hashes all fields which are taken into account are normalized, i.e. certain special characters are removed, whitespace and author/editor names normalized. The latter is done by concatenating the first letter of the first name by a dot with the last name, both in lower case. Persons are then sorted alphabetically by this string and concatenated by a colon.

The current duplicate detection is very simple and fails to detect spelling errors, differences in how special characters (like German umlauts) are entered or additional LATEX commands. This is ongoing work; our implementation allows for simple addition of new hashes.

Currently, duplicate detection is used on the one hand to warn the user when she wants to add an already existing resource and on the other hand to show how many users tagged a certain resource. A step beyond detecting duplicates could be providing the user with additional fields found in other entries referring to the same publication so that she can complete her own entries with additional information.

For bookmark entries in BibSonomy, their URLs are currently just hashed with MD5 and this hash is used for comparison. As can be seen from discussion with users, opinions on if and how to normalize URLs in such systems differ. On the one hand, URLs like `http://www.w3c.org`, `http://w3c.org` and `http://www.w3c.org/index.html` might denote the same resource, on the other hand they’re different URLs and it is not obvious whether they really mean the same resource.

Editing tags Besides changing the tags of a post by editing it, BibSonomy offers at the moment two other ways of changing the tags of several posts at once.

By preceding the path part of a personal URL (i.e. one where the path starts with /user followed by the users own username) with /bediturl (or /beditbib) one can edit the tags of all bookmarks (or publications) on the page at once. This function is also available through links on the respective pages.

Furthermore we have an $m:n$ tag editor which allows a user to exchange m tags by n other tags. More precisely: given two sets A and B of tags[11] and a user $u \in U$, then the $m:n$ tag editor sets

[11] If B contains tags not already included in T, then T is adjusted in the obvious way.

iteratively for every $r \in R_u$ with $A \subseteq \text{tags}u, r$:

$$Y := (Y \setminus (\{u\} \times A \times \{r\})) \cup (\{u\} \times B \times \{r\}) .$$

Both functions support the user in creating and maintaining a consistent tag vocabulary.

Import of resources To encourage users to transition from other systems we implemented an import functionality. For del.icio.us, this functionality also takes into account the del.icio.us *bundles*. They are mapped to BibSonomy's relation $\prec$ in the following way: for every bundle B (which is a set of tags) with name b we add $\{b\} \cup B$ to T and set

$$\prec := \prec \cup (\{u\} \times B \times \{b\})$$

where $u \in U$ is the user these bundles belong to. Furthermore it is possible to import bookmark files of the Firefox[12] web browser, where the typical folder hierarchy of the bookmarks can be added to the users $\prec$ relation. That means that, for every folder a and every subfolder b of a in Firefox, we add (u, b, a) to the users u is-a relation $\prec$, if the user chooses to do so.

Import of existing BibTEX files is also simple: after uploading the file, the user can tag the entries or automatically assign them the tag *imported*. If a BibTEX entry contains a field *keywords* or *tags*, its contents are attached as tags to the resource and added to the system. BibTEX-Fields unknown to BibSonomy are saved in the *misc* field and will not get lost.

Export of resources Exporting BibTEX is accomplished by preceding the path of an URL with the string /bib – this returns all publications shown on the respective page in BibTEX format. For example the page `http://www.bibsonomy.org/bib/search/text+clustering` returns a BibTEX file containing all literature references which contain the words “text" and “clustering" in their fulltext.

More general, every page which shows posts can be represented in several different ways by preceding the path part of the URL with the string described here:

/ the typical HTML-View with navigation elements

/xml bookmarks in XML format

/rss bookmarks as RSS feed

/bib publications in BibTEX format

/endnote publications in EndNote[13] format

/publ publications in a format suited for integration into a homepage (for an integration example see `http://www.kde.cs.uni-kassel.de/schmitz/publikationen.html`)

/publrss publications as RSS feed

For example, the URL `http://www.bibsonomy.org/publrss/tag/fca` represents an RSS feed showing the last 20 publications tagged with the tag *fca*.

These export options simplify the interaction of BibSonomy with other systems. RSS feeds allow easy integration of resource lists into web sites or RSS aggregators and BibTEX output can be used to automatically generate publication lists for papers (as done with this paper). With the help of XSLT it is also possible to transform the RSS output into formats suitable for further processing. In addition further formats are implemented easily by extending the URL scheme and adding an appropriate JSP which generates the output. Further formats have been implemented, based on the JabRef system.

[12] `http://www.mozilla.com/firefox/`
[13] `http://www.endnote.com`

Groups In many situations it is desirable to share resources only among certain people. If the resources can be public, then one could agree to tag them with a special tag and use that tag to find the shared resources. The disadvantage is, that this could be undermined by other users (or spammers) by using the same tag. To solve this problem and also to allow resources to be visible only for certain users, we introduced *groups* in BibSonomy which gives users more options to decide with whom they share their resources.

It is thus possible to have private posts, which only the user can see, as well as posts which can be seen only by group members. Overall there are several aspects of groups in BibSonomy:

1. One can get an aggregated view of all resources of the group members. For example, the URL `http://www.bibsonomy.org/group/kde/seminar2006` represents all posts the members of the *kde* group have tagged with *seminar2006*.

2. It is possible to use groups for privacy so that certain references can only be seen by group members.

3. Resources can be copied directly to the group so that they're persistent, even if a user leaves the group. This is possible, since groups are implemented as a special user which has the name of the group and owns the copied references; this user is also the group admin. This feature is in particular useful where the donator has to commit to the resource, e.g. for project deliverables or student projects.

While we are using one group for sharing resources within our institute, we run several other groups which are used for teaching – they collect resources for students which they might find useful for their lecture.

Shopping Basket Every publication can be "picked" and is then available in a "shopping basket"-like download area. This is useful for collecting references one needs for a publication. Since all publication related export options mentioned in Section 2.2.2 apply to this, it is straightforward to get all collected posts in BibTEX or EndNote format.

BibTEX Styles There exists a vast amount of different styles[14] for BibTeX and it is tempting to use them for generating nicely formatted HTML output. Different outputs have recently been implemented using JabRef.

API Experience has shown, that an Application Programming Interface (API) is crucial for a system to gain success. It is something which has been requested by many people and which allows for easy interaction of BibSonomy with other systems. Hence we are currently working on the implementation of a REST API (Fielding, 2000) which can be used and accessed also by not so experienced programmers.

Information Extraction for publication references import With the first version of BibSonomy, literature references could be imported only from proper BibTEX source code. This was a strong restriction, since most literature references in the web are not available in BibTEX format but rather in the form of human readable publication lists. Hence our efforts to enhance import focus on techniques to allow for the import of such resources. We have recently implemented information extraction techniques (Peng and McCallum, 2004) from the MALLET system (McCallum, 2002).

[14] `http://www.cs.stir.ac.uk/~kjt/software/latex/showbst.html`

Further BibSonomy functionality Additionally, beside the features listed above, the following extended functionalities have been added to BibSonomy: keyboard shortcut for BibSonomy posting in firefox, OpenURL support, tag editor, OWL output, tag hierarchy, gnome desktop integration, scrapers for ACM Digital Library and CiteSeer, spam filter, logging of copy button, improvement of basket and group functionalities, tutorials, faq, extended help pages, migration to a new server to increase hardware redundancy, password forgotten functionality, improved relation management, information extraction for publications in unstructured text, customizable export formats (including CSV for spreadsheets, HTML, RTF for Word and other text processors, DocBook XML), fulltext search. For more details see `http://bibsonomy.blogspot.com/` and Deliverable 2.1.

Task 2.1.2: Folsonomy peer-to-peer system for sharing of bibliographic data

The prototype distributed tagging system has been developed in Java to allow its use on virtually any platform. The system is based on the Semantic Exchange Architecture (SEA) (Franz et al., 2006) - initial work which defines some essential components for a scalable folksonomy peer-to-peer architecture. Internally a tagging ontology is used to represent all tagging data as RDF graph and each peer maintains a metadata repository to store this tagging information.

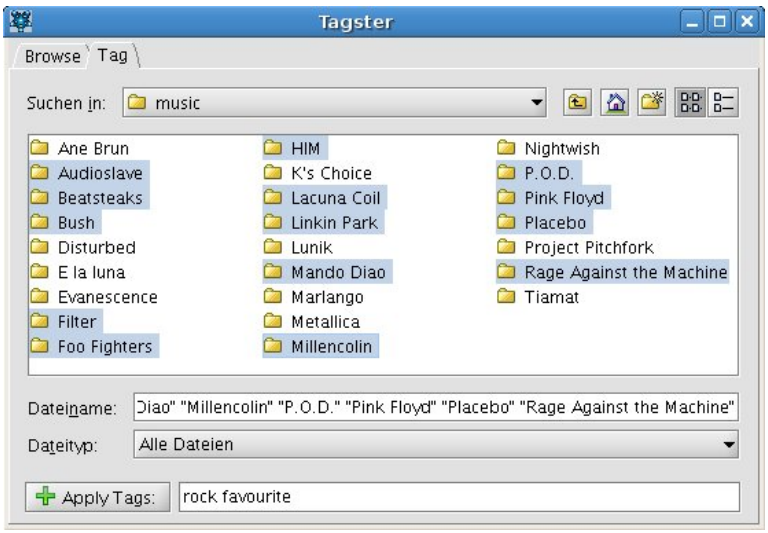


Figure 2.3: Peer-to-peer application - file browser like view for tagging data.

The SEA kernel is the core module with direct access to the repository and the network interfaces. The modularization also distinguished between server and client components which makes it basically possible to have a single repository for a small working group with several connected user sharing the data.

For the interaction with other peers two different network layers are necessary. Both have been already implemented with some basic functionality to exchange data and metadata between the peers. The first network layer is responsible for managing the direct interaction with known peers and trusted buddies to directly retrieve the tagging data and information objects. A basic tag-based trust mechanisms ensures that a peer can only see the data that was explicitly made available. Thus it is possible to differentiate between public and private data. For disclosed information objects all other peers can see the visible tags and can also add some own tags.

Additionally, some multimedia data inspection features have been implemented to automatically extract specific multimedia metadata which can be directly added as tags. This helps to simplify the tagging process - especially for larger data sets like music collections. Through the inclusion of additional ontologies it is easy to semantically enriched the system, e.g. qualifying that certain metadata represents author, title, bitrate etc. of some media file.

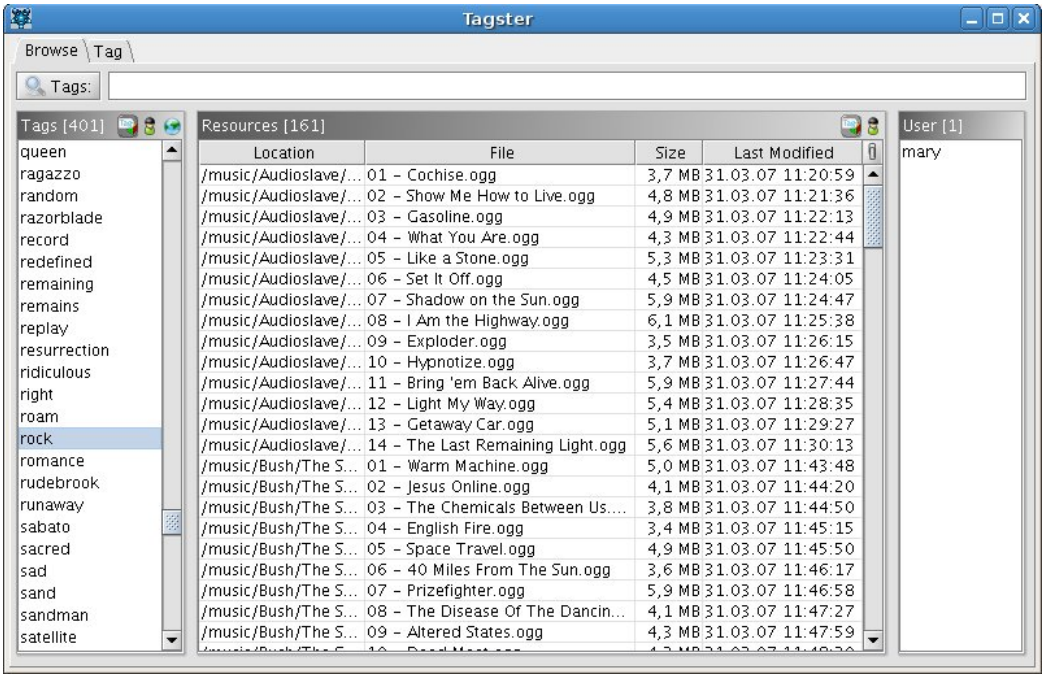


Figure 2.4: Peer-to-peer application - showing all tags and all resources tagged with “rock”.

The system also needs to present the user some useful statistics about the tagging data (like similar tags etc.) which helps to navigate the tags and leads to new interesting information. The second network layer is using a distributed index structure to make such information globally available within the network. The prototype stores for example user-tag, user-resource, and resource-tag relations. Thus accessing a resource ID in the index will return all tags assigned to that resource and all users who tagged it. The current index structure has been implemented on top of a distributed hashtable which makes single lookup operations very efficient.

However, for complex tagging data analysis like co-occurrence and similarity computations such a simple index structure is not very efficient because multiple index lookups would be necessary. To find a more sophisticated solutions for this problem we have been investigating what kind of folksonomy measures are really important for common tagging statistics and which of those would be efficient to be implement in a distributed scenario. We started form a general analysis of the complexity of similarity computation (as for clustering) and how such measures can be approximated to reduce the overall computation overhead when working with a large tagging dataset. Secondly, have been looking for suitable distributed data structures that can be efficiently implemented in our peer-to-peer scenario.

The basic idea for our similarity measure is derived from *term frequency* and *document frequency* as used in information retrieval and can (with some adjustments) be applied to tagging data as well. With this approach certain approximations allow us to reduce the overall message overhead required to update the tagging statistics stored in the network. Details of this work will be made available in an upcoming publication also including formal description of a generic similarity measure for tagging data.

In the future the distributed tagging system should be available as a browser plug-in to minimize the user effort concerning installation and updates. Therefore, it has been investigated which platform would be best suited and what requirements are necessary to ensure that many people can use the software even when connecting from within secured subnets. Some techniques for detecting and circumventing firewalls have been explored and shall be used in future versions.

Task 2.2 Tag-based navigation systems

Over the past year, Sony CSL developed the Ikoru system, which is primarily used to experiment with collaborative tagging and content-based analysis. The current web interface, viewable at `http://www.ikoru.net`, handles both image and music files. A demo version with testing data is available at `http://demo.ikoru.net`.

Ikoru's web interface resembles those of existing tagging sites. However, some of the features of Ikoru make it unique:

Content-based tools: The initial motivation of the Ikoru project was to explore to combination of content-based analysis with tagging (see also section 2.3.1).

Data Gathering: Ikoru is used as a platform to gather data and explore how the analysis of this data can improve tagging systems. Not only do we store tagging information but also the history of the visitors (visited pages, performed searches). This data may improve analysis techniques that currently only take tagging information in account.

Extendible research platform: Ikoru aims to be an open platform that can be extended with new analysis and visualization tools. As such, researchers can integrate the analysis results in the web site and evaluate the impact of the these tools on the behavior of the users.

Multimedia: Although initially developed for images, the system has been extended to handle music files as well. Future versions will also include video files and text. This multimedia aspect is not incidental. We want to study the tagging behavior for various media types and we seek to use the semantic layer created by the tags as a support for new creative tools.

Small, reusable server: On the technological side, we have designed the Ikoru server as a small, efficient, and stand-alone web component that can be easily reused and integrated in third-party projects.

The similarity-based search tool discussed in section 2.3.1 was integrated in the web site. By drag-and-dropping a photo icon on the search tool, the photos of the context will be sorted according to the visual similarity with the example image. Up to three positive and negative example images can be given. The similarity measure can be based on color, texture, or both. To enable the similarity search, the Ikoru server computes two features vectors when a photo is uploaded: one for color and one for texture.

We added has a server-side scripting interface to Ikoru that allows third party to execute analysis programs on the server in a similar way to the Common Gateway Interface (CGI). The advantages of the open platform are that it speeds up the testing of a hypothesis, that the analysis can be applied to user-defined contexts, and that the results can be easily shared with others. Finally, the results can be integrated in the web interface to study its impact on the behavior of the users.

The Ikoru system was initially developed for images but handles music files, too. To enter the editorial meta-data to the system, such as the artist name or the album title, to music we choose to use "machine tags" instead of using a more elaborate database schema as is common practice. Machine tags were introduced to annotate, for example, geographical information to photos. Machine tags have the following syntax: $namespace : predicate = value$. We subsequently adapted the existing meta-data to this format. For example, "music:artist=the_beatles" indicates that the music file is a recording of "The Beatles". Using this approach we can evaluate whether tags are a viable method for storing conventional meta-data.

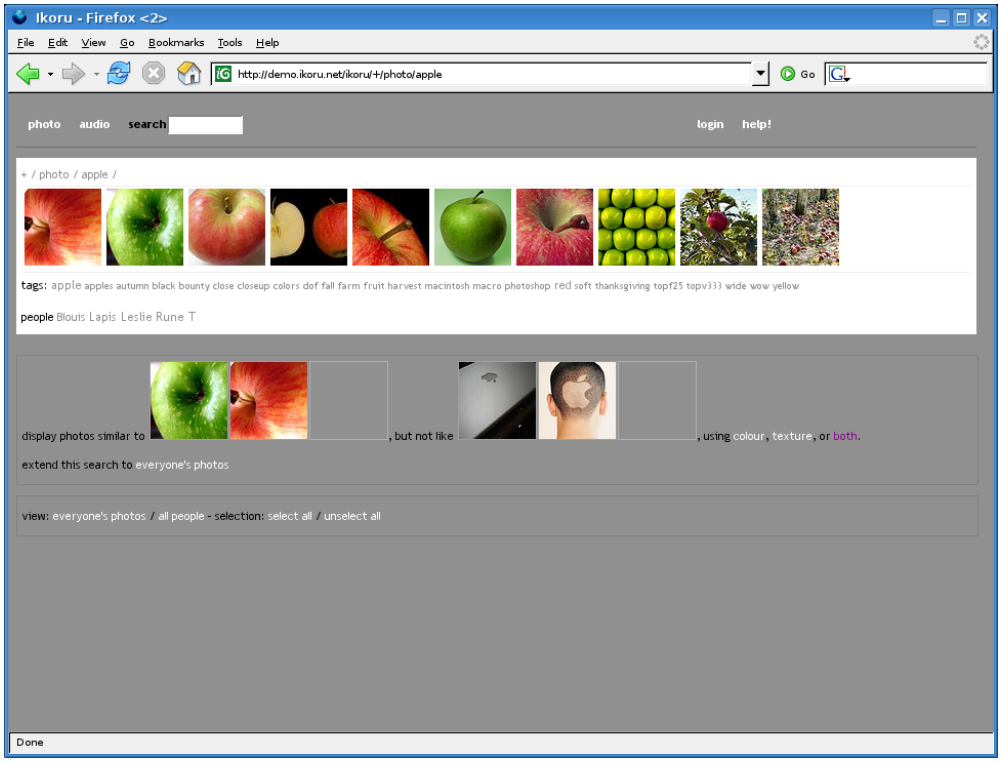


Figure 2.5: A screenshot of Ikoru showing the similarity-based search. showing the similarity-based search. showing the similarity-based search.

The Ikoru system consists of two components: a server and a web interface. The server was developed as a stand-alone application written in C++ (40.000 source lines of code). It publishes its Application Programming Interface (API) on the Internet using the Web Service Description Language (WSDL). Third-party client libraries can use the WSDL description to dynamically bind to Ikoru and exchange information using the SOAP protocol. Client libraries for JavaScript, Java, and PHP are available, in addition to a partially implemented C++ client library.

The server Application Programming Interface (API) uses the *context* object as a central pillar. A context is uniquely identified by a string that resembles a file path. It is composed of a user name, a media type, zero or more tags, and, optionally, a resource identifier. For example, the identifier "/hanappe/photo/tagora/2007" comprised all the photos of user "hanappe" tagged with "tagora" and "2007". There is a direct mapping between the context identifier and the URL of the Web page to view the context.

The web interface was implemented using the AJAX paradigm. It was developed to make our work accessible online and attract a community of users. It has gone online recently and will start promoting it in the coming year.

2.2.3 Milestones

M2.1 (Task 2.1.1) First version of social tagging system for bibliographic data (month 5).

Until the milestone, the following extended functionalities (see above) have been added to BibSonomy: keyboard shortcut for BibSonomy posting in firefox, OpenURL support, tag editor, OWL output, tag hierarchy, gnome desktop integration, scrapers for ACM Digital Library and CiteSeer, spam filter. For more details see `http://bibsonomy.blogspot.com/` and Deliverable 2.1.

2.2.4 Deviations and Corrective Actions

SONY-CSL The initial project proposal mentioned two subtasks for Task 2.2: first, a tag-based navigation systems for images, and second, a tag-based navigation systems for music. These two systems have been integrated in the Ikoru platform. This integration is more than justified by the gain in effort and the reuse of existing work. In addition, it opens up new opportunities to study and exploit tagging information for different media types.

UNI KO-LD *Task 2.1.2* The peer-to-peer tagging system was inspired by Bibster (Haase et al., 2004), a peer-to-peer sharing system for bibliographic data. But due to the limitations of Bibster which does not allow arbitrary tags and does not support statistics a new distributed tagging infrastructure has been developed. It is based on the Semantic Exchange Architecture (SEA) (Franz et al., 2006) and allows for tagging arbitrary information objects. The use of a separate network layer based on Bamboo[15], a distributed hashtable implementation, provides an easy way to distribute simple tagging statistics in the network.

With the new architecture and the possibility to tag any type of resource we changed the focus from bibliographic data to multimedia data for three main reasons. First, we avoid a direct competition with BibSonomy. Second, we expect to attract more users interested in a tagging multimedia data than bibliographic date. Finally, a much larger set of resources and therefore tagging data will be available because the storage of multimedia data on personal computers is much more common than bibliographic data. Consequently, JabRef is not suitable as a client any more. Thus we have been developing a new client application that allows for easy tagging, organization, and sharing of the multimedia data. It is coded in Java as well to allow its use on virtually any platform.

According to the original work we had an initial delay of about 2 month due to hiring of additional staff after the project started. However we tried to cope with that delay through additionally hired junior staff and the contributions of other non payed TAGora workers. UNI KO-LD globally used more man-months on WP2 and less man-months on WP4 since the first year has been mainly devoted to the developing of the applications. The total number of man-months used is as planned.

[15] `http://bamboo-dht.org`

2.2.5 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
2.1	Task 2.1a First version of social tagging system for bibliographic data.	2	31 May 2007	31 May 2007	5	4	**UNIK**
	Task 2.1b First version of folksonomy peer-to-peer system for sharing of bibliographic data.	2	31 May 2007	31 May 2007	5	7.6	**UNI KO-LD**
2.2	Task 2.2a First version of the Tag-based navigation system for images.	2	31 May 2007	31 May 2007	3.5	3.5	**SONY-CSL**
	Task 2.2b First version of the Tag-based navigation system for music.	2	31 May 2007	31 May 2007	3.5	3.53	**SONY-CSL**

Mil. No.	Milestone name	WP No.	Date due	Actual/Forecast delivery date	Lead contractor
M2.1	First version of social tagging system for bibliographic data.	2	31 October 2006	31 October 2006	**UNIK**

2.3 Workpackage 3 (WP3) - Data analysis of emergent properties

2.3.1 Objectives

Following are the objectives of the research carried out during the first year of the project.

In WP3 the data sets coming from WP1 should be analyzed for general features common to the different systems in study and also for specific features which help to discriminate them. This includes the identification of suitable measures which can be performed on the datasets as well as acquiring existing and developing new software tools for performing the measures. Once the measures and tools are collected they will be applied on the datasets. There is an important interaction with WP4 where models for describing collaborative systems are developed where the results from the data analysis will be used for verifying whether models apply.

Task 3.1 Emergent metadata statistics

Task 3.1 deals with the quantitative statistics coming from the analysis of the datasets provided by WP1. This includes basic information like the number of users, tags resources and tag assignments but also frequency distributions of tags and number of tag assignments per user. One goal is to investigate whether all the data sets expose similar power law distributions which is commonly referred to as the long tail and expresses the situation that for example few users are frequent

taggers while the majority of users just assigns a few tags to few resources.

Another important aspect is the evolution over time. This considers the increase of the number of users, tags (i.e. the vocabulary), and resources which may also adhere to a certain distribution and can indicate whether a folksonomy converges on terms and fosters consensus between the users. Consequently if the distribution flattens or narrows it perhaps means less or more agreement. This kind of analysis, borrowing concept and methods from Probability Theory and Time Series Analysis, will benefit and orient the modeling of systems in terms of Naming Games Model and Stochastic Models (see WP4).

Task 3.2 Network/graph analysis

The network/graph structure of folksonomies encloses the whole complex relational, semantic, social information accumulated by the on-line social community. A detailed analysis is crucial to fully characterize the structure, starting from its simplest topological features. Such an analysis has a twofold aim. Firstly, it is essential for the theoretical description of the system: it allows for a quantitative comparison between different folksonomies, hence suggesting the genuine universal features a model should capture and describe. Secondly, the analysis could evidence weakness and vulnerability of the systems, forecast its scaling trends and drive every future control activity and interface improvement.

The study of complex network witnessed a surge of interest during the last few years in the complex science community. It has become a successful paradigm useful to frame a variety of problems and system study, most of them resulting from several human and social activity. A number of concepts, measures, tools and models have been proposed in order to capture the essence of the rich phenomenology observed in real systems.

However many of these results are mainly suited to simple mono-partite (directed or undirected) networks, and have to be explicitly recasted, adapted or extended, for the case of tri-partite structure typical of folksonomy. Alternatively, specific projection methods have to be considered in order to reduce the full structure in a more comfortable and usable form.

The first year of the project is mainly devoted to gather, to perform first global statistical analysis, and to settle the computational facilities needed for the full network analysis. However, a first research activity can start on partial snapshots of the dataset, in order to identify and test the theoretical tools needed for the future.

Task 3.3 Cluster/community identification

A more advanced study of the network structure of folksonomies should certainly attack the problem of cluster and community detection: for folksonomies, this is an highly non trivial task, due to their scale-free properties. At the same time, obtaining results in this direction is probably one of the most important goal of the project, since it can provide a method to extract the relevant information emerging from the social activity, improve knowledge navigation and data mining. For instance, a cluster identification applied to the resources posted by the users, could provide a effective and socially agreed classification of topics and knowledge. On the other hand, one could in principle identify cluster of “connected” tags in order to extract a classification scheme, or even a hierarchy of concepts, a shared and consistent auto-organized semantics. Finally a successful community detection strategy for the users could be useful to implement recommendation strategies.

The theoretical and experimental activity in cluster and community detection should be concentrated in the last two years of the project: however some preliminary experiments could be performed as soon as data were available.

Task 3.4 Semantic inference

Pure quantitative statistics of folksonomy data is not enough to fully understand the user behavior and derive suitable models. For a comprehensive view on folksonomies it is necessary to include background knowledge because users usually do not apply different tags independently. In contrast, conceptually related tags like “cat” and “animal” are often used together. Co-occurrence analysis helps to identify some of those relation but also has its limitations, e.g. when it comes to language boundaries. Ontologies have gained a lot of prominence for knowledge sharing, but have so far found little entry into such data analysis. Therefore the inclusion of ontological knowledge is expected to significantly improve the identification of data relations in respect to the granularity of the targeted analysis as well as the granularity of ontology concepts. We want to investigate new methods considering such background information to find hidden information or better explore the relations in sparse datasets.

Task 3.5 Cross-Folksonomy Networks

Folksonomy web sites are rarely closed worlds. It is quite common for individuals to be active members of several online communities and thus one would expect certain tags to spread across such communities with time. For example one could be adding images to Flickr, bookmarking web sites with del.icio.us, creating their music preference profiles in last.fm, and tagging articles in Connotea. By continuously collecting data from such folksonomy web sites, and monitoring changes and additions, we can cross reference emerging tags between the separate communities to extend and connect their individual networks to create an Integrated Semantic Network. Tags might be the only common parameter between the various folksonomies and therefore it can be used to link the separate networks together. Social Network Analysis and Graph Theory can be applied to this overall network to, for example, recognize cross-folksonomy communities that used certain tags to describe various types of objects. Such analysis can help to identify clusters of objects that are of different kinds (e.g. an image and a document and some music in one cluster), but were given the same or related tags. Such non-homogeneous clusters are very difficult to identify when analyzing folksonomy web sites in isolation due to their tendency to deal with specific type of objects. This task will therefore ensure that some of the separate folksonomy networks that will harvested within this project are connected together.

Once the networks are integrated, we will apply Network Analysis and Visualization techniques to study how tagging evolved and spread across the various communities (Alani et al., 2005), and how the commonalities between these communities can be employed as a basis for recommendations that could go beyond what most of these folksonomies currently provide. Part of this work will focus on the design and implementation of system architecture to link some of the folksonomy networks that will be provided by the consortium.

Task 3.6 Collaborative tagging and emergent semantics

This task was not envisioned in the original plan of the project, and has been created for the reasons explained in the deviations section. The objective of this task is to study the relation between tags and content-based analysis. Is it possible to *ground* tags? Is it possible to improve the navigation based on tags with data extracted from the content? Can we reduce some of the limitations of tagging, such as the problems of homonymy and synonymy? And vice versa, can tags offer a support for automatic classification schemes? These are some of the questions that we aim to address.

2.3.2 Progress

Task 3.1 Emergent metadata statistics

According to the above mentioned objectives, different quantitative statistics were analyzed in the data sets provided by WP1. The largest datasets are those from del.icio.us (667,127 users; 2,454,546 tags (organized in 667 bundles); 18,782,132 resources; 140,333,714 tag assignments) and flickr (298,954 users; 1,553,253 tags; 24,599,875 resources; 110,345,103 tag assignments).

A typical folksonomy measure is the probability distribution of tags which has also been investigated for the delicious and BibSonomy datasets. The analysis for the two datasets clearly identified specific anomalies like spam postings. After removing the spam postings from both systems one can observe very similar distributions which means that they expose the same features.

Another important aspect is the stream analysis that considers the changes over time. For example, the relative fraction of the cumulated tag occurrences was investigated as a function of the age of a resource or user, i.e. the age of a resource or user will be measured in the number of postings assigned to a resource or assigned by a user. Typically, the time is measured in number of postings and not in e.g. days because the popularity of a resource has an important influence on how fast the tag fractions reach a stable state. The results show that the relative proportions of the most popular tags at a resource reach a quite stable state after an initial transient. In some cases it is also possible that new tags overtake already established tags. Quite interesting in addition to that is that the growth of the used vocabulary (i.e. tags) is exponential. This behavior is already known form studies on text corpora and has now also been observed in different streams of delicious data.

Furthermore, we explored in how far techniques known from record linkage can be used for merging different spellings of tags based on their string similarity. A testing tool has been implemented that uses a freely available Java library with several string comparators (Levenshtein, Jaro-Winkler, ...). It may also be possible to merge e.g. different flexions of tags like the singular and plural form of nouns. We performed preliminary experiments on tagging data obtained from Delicious in which we tested whether merging tags based on their string similarity is suitable for cleaning up the raw tagging data.

We also analyzed in how far the users show different behavior with regard to handling flexions of nouns, i.e. whether they preferably use singular or plural, and with regard to compound nouns. Of special interest was, whether one can observe a change in the user's behavior over time. The results show that most users apply the singular form more often than the plural form. But there is also a certain amount of users which often uses both forms. The ratio between singular and plural forms were constant over the whole time. In contrast for the compound words, where a change in the ratio between the different forms of handling them could be observed. For both experiments, we used Wordnet as a lexical resource for linguistically analyzing the tagging data and for concluding on the ongoing semiotic dynamics in tagging systems.

Task 3.2 Network/graph analysis

The structure of collaborative tagging networks has been analyzed by adapting characteristic path length and clustering coefficients to the three-partite tagging data. Such measures have been introduced in order to manifest and quantify the *small world* property of complex networks (Schmitz et al., 2007).

The characteristic path length of a graph (Watts, 1999) describes the average length of a shortest path between two random nodes in the graph. If the characteristic path length is small, few hops will be necessary, on average, to get from a particular node in the graph to any other node. In the case of a triadic structures of *(tag, user, resource)* assignments, the effort of getting from one node in the folksonomy to another can be measured by counting the *hyper* edges in shortest paths

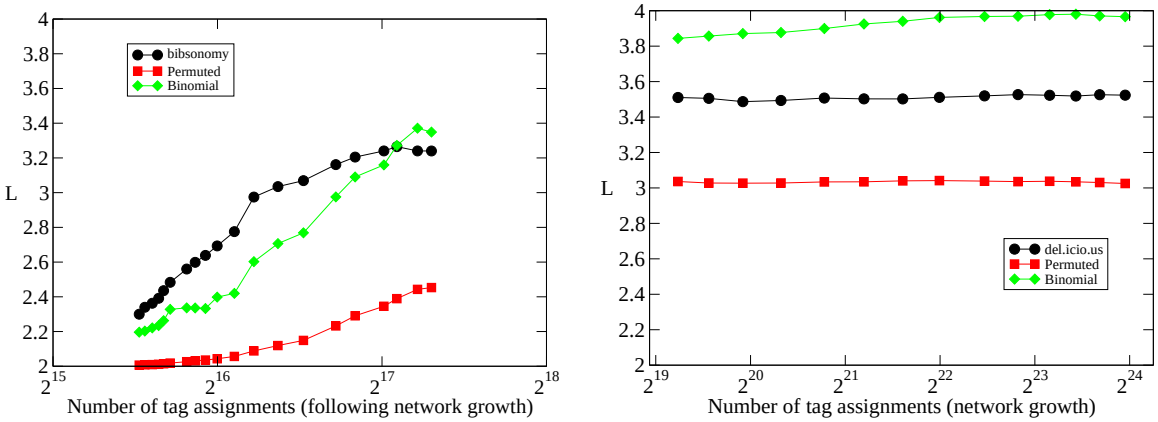


Figure 2.6: Characteristic path length for the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graph have been obtained as a function of the number of nodes of the networks (not shown). Note how the characteristic path length takes quite similar low values, typical of small world networks, for all graphs.

between the two.

In Fig. 2.6, measures of the characteristic path length in folksonomies are shown. As can be seen the small values observed reveal the *small world* nature of the networks. Interestingly the results in the two folksonomies are comparable, in the sense that the values measured on younger, smallest system `BibSonomy` seems to converge to the almost constant values of `del.icio.us`, considered as a mature system. In both cases we compare the results with two kinds of derived random graphs:

Binomial: These graphs are generated similar to an Erdős random graph $G(n, M)$ (Bollobas, 2001). T, U, R are taken from the observed folksonomies. $|Y|$ many hyperedges are then created by picking the three endpoints of each edge from uniform distributions over T, U, and R, resp.

Permuted: These graphs are created by using T, U, R from the observed folksonomy. The tagging relation Y is created by taking the TAS from the original graph and permuting each dimension of Y independently (using a Knuth Shuffle (Knuth, 1981)), thus creating a random graph with the same degree sequence as the observed folksonomy.

The case of clustering coefficient is more subtle. Clustering or transitivity in a network aim to measure how much the network is locally dense around each node. In the case of three-mode data this definition combines two aspects which are *not* equivalent:

Cliquishness: This "clustering coefficient" of a node is high iff many of the possible edges in its neighborhood are present: i.e. considering a resource r and the set of tags T_r and users U_r connected to r, then the neighborhood of r was maximally cliquish, if all of the pairs from $T_r \times U_r$ would occur in hyper links (TAS) of the graph. The same definition of cliquishness stated here for resources can be made symmetrically for tags and users.

Connectedness (Transitivity): The other point of view follows the notion that the clustering around a node is high iff many nodes in the neighborhood of the node were connected even if that node was not present.

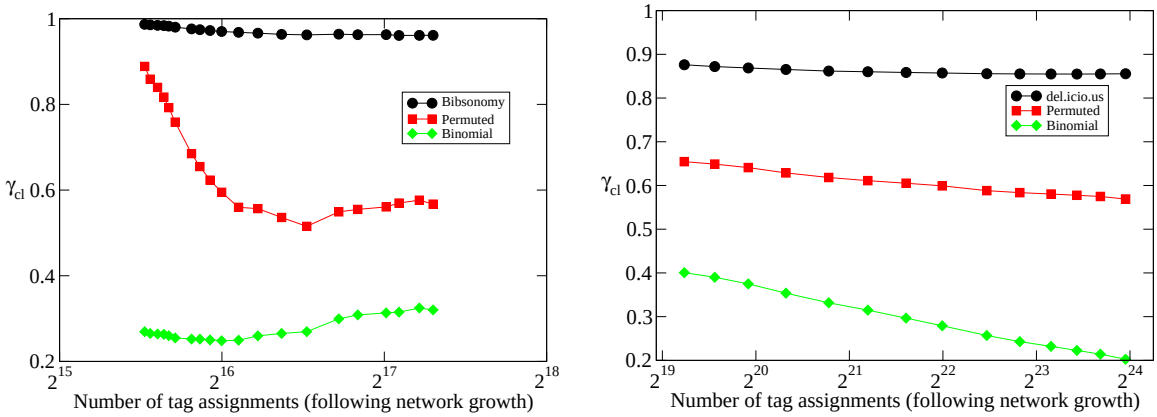


Figure 2.7: Cliquishness of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). The cliquishness for the folksonomy networks takes quite high values, higher than the corresponding random graph (permuted and binomial).

These measures performed on *BibSonomy* and `del.icio.us` are shown in Fig. 2.7 and Fig. 2.8, respectively for cliquishness and connectedness. Both measures are compared with the randomized graphs introduced above.

Again, all these measures confirm that folksonomies possess the *small world* properties. Moreover they give comparable results in the two folksonomies considered, suggesting a general universal behavior, spontaneously emerging from the activity of users. The last observation is very important, since it support the use of these quantities as good descriptors of the network topology.

On the other hand, one could use more standard measures, provided a suitable projection of the folksonomy tripartite graph on a standard graph is chosen. An example is reported in (Schmitz et al., 2007). There we project the folksonomy hypergraph on an undirected, weighted network of tags: the weight between two nodes (tags) being the number of posts the two tags co-occur.

The reason to define such a graph is to analyze it with some common statistical tools:

- The cumulative strength distribution, i.e. the distribution of node strengths, defined as the sum of weights of links connected to the node;
- Average nearest neighbor strength, or node-nearest neighbor strength correlation, which measure for each node the average strength of its neighbor nodes.

.

We measured these quantities in both `del.icio.us` and BibSonomy snapshot. The first observation is that the statistical properties for the two systems are very similar and similar to other scale-free complex networks. typical of social networks. Interestingly, some statistically appreciable differences have been easily recognized to be due to few spammer users.

Furthermore, we adopted a shuffling protocol on the tags in the posts, similar to the one used above to generate the permuted random graph, with the aim to explicitly study the role of tag semantics.

In Fig. 2.9 we show the cumulative strength distribution for the network of co-occurrence of tags in `del.icio.us`. The two kinks pointed by arrows disappear after the network have been filtered by spammers contributions, leaving a very clean power law shape (similar results have been obtained for BibSonomy). However the shuffling procedure, which destroy all semantics in the network,

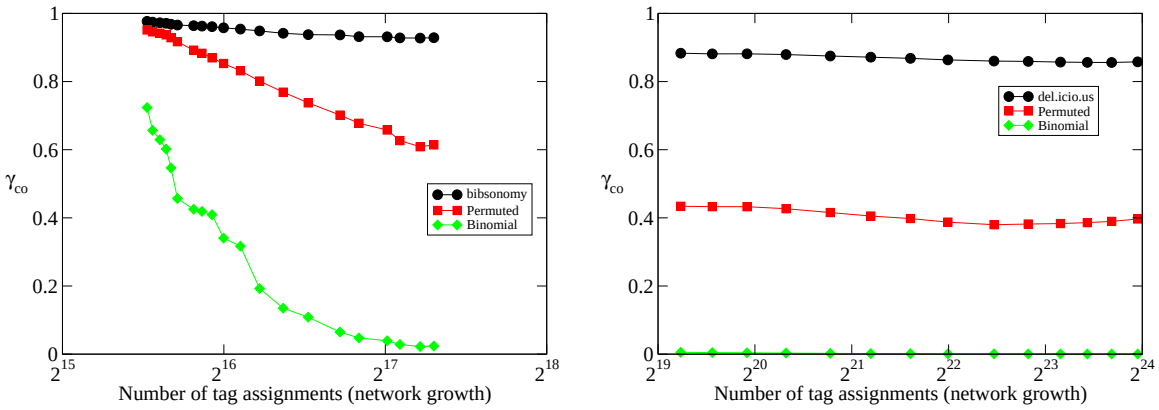


Figure 2.8: Connectedness/Transitivity of the `BibSonomy` folksonomy (left) and the `del.icio.us` folksonomy (right), compared with the corresponding random graphs: permuted and binomial (see text). The measure is repeated following the network growth and shown as a function of the number of tagging events. Similar graphs have been obtained as a function of the number of nodes of the networks (not shown). As in the case of cliquishness, the values of connectedness/transitivity are very high for the folksonomy networks, at odds with the corresponding random graphs (permuted and binomial).

only slightly affect the measure, pointing that its scale-free shape is only due to the frequency distribution of tags, similarly to the Zipf's law observed in texts.

On the contrary, the average nearest-neighbor strength, plotted in the scatter plot of Fig.2.10 (data from `del.icio.us` snapshot), presents a much richer structure. Again anomalous network nodes due to spammers, manifest themselves as disjointed clusters of points, and can be easily removed from the analysis. In this case, the shuffling of tags affects the plot quite strongly. The analysis suggests that many "assortative" nodes of the network (i.e. weak nodes/tags co-occurring with strong nodes/tags) carry important semantic information, this pointing to an underlying hierarchical organization of tags (similar results have been observed for BibSonomy).

An additional activity, orginally not planned in the Annex I, concerned the efforts devoted to the adaptation of successfully ranking algorithm (PageRank) to the specific structure of the tri-partite networks of folksonomies. This work has been done in cooperation with the EU FP6 IP *Nepomuk – Social Semantic Desktop*, and has been published as (Hotho et al., 2006c). Our approach is based on our *FolkRank* algorithm (Hotho et al., 2006b), a differential adaptation of the PageRank algorithm (Brin and Page, 1998) to the tri-partite hypergraph structure of a folksonomy. Compared to pure co-occurrence counting, FolkRank takes also into account elements that are related to the focus of interest with respect to the underlying graph/folksonomy. In particular, FolkRank ranks synonyms higher, which usually do not occur in the same bookmark posting together.

With FolkRank, we compute topic-specific rankings on users, tags, and resources. In a second step, we can then compare these rankings for snapshots of the system at different points in time. We can discover both the absolute rankings (who is in the Top Ten?) and winners and losers (who rose/fell most?).

The contributions of this work are:

Ranking in folksonomies We describe a general ranking scheme for folksonomy data. The scheme allows in particular for topic-specific ranking.

Trend detection We introduce a trend detection measure which allows to determine which tags, users, or resources have been gaining or losing in popularity in a given time interval. Again,

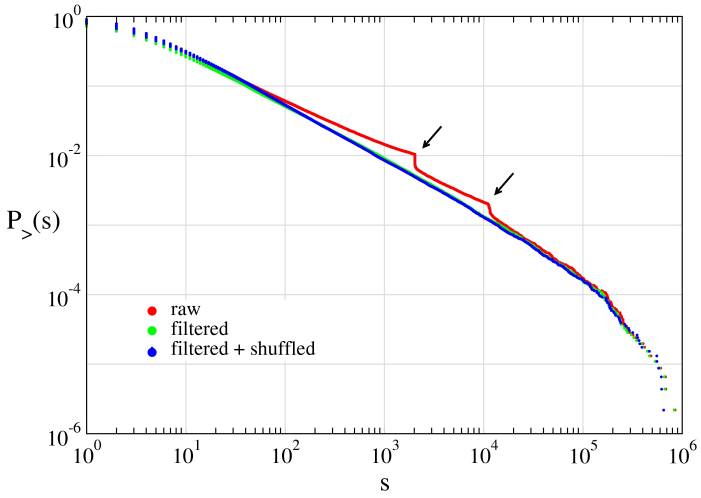


Figure 2.9: Cumulative strength distribution for the network of co-occurrence of tags in *del.icio.us*. $P_>(s)$ is the probability of having a node with strength in excess of s. Red dots correspond to the whole co-occurrence network. The two steps indicated by arrows correspond to an excess of link with a specific weight and can be related to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps (green dots). Shuffling the tags contained in posts (blue dots) does not affect significantly the cumulated weight distribution. This proves that such a distribution is uniquely determined by tag frequencies within the folksonomy, and not by the semantics of co-occurrence.

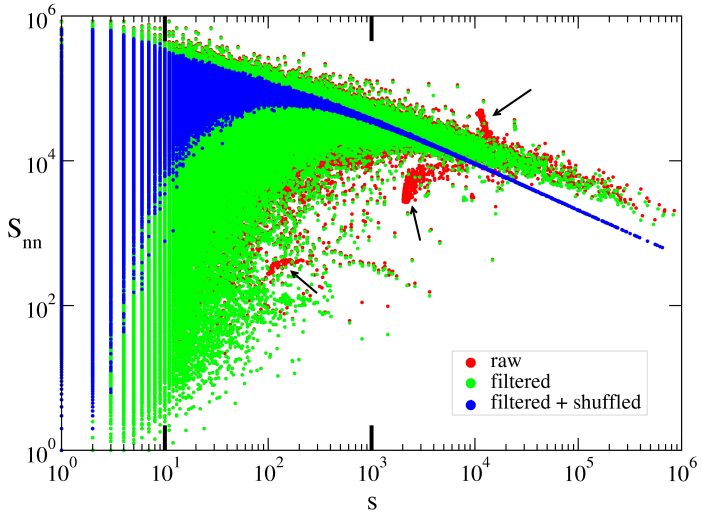


Figure 2.10: Average nearest-neighbor strength S_{nn} of nodes (tags) as a function of the node (tag) strength s, in *del.icio.us*. Red dots correspond to the whole co-occurrence network. Assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. A few clusters (indicated by arrows) stand out from the main cloud of data points. As in Fig. 2.9, such anomalies correspond to spamming activity and can be removed by filtering out posts containing an excessive number of tags (green dots). In this case, shuffling the tags (blue points) affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the network of co-occurrence beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics.

this measure allows to focus on specific topics.

Application to arbitrary folksonomy data As the ranking is solely based on the graph structure of the folksonomy – which is resource-independent – we can also apply it to any kind of resources, including in particular multimedia objects, but also office documents which typically do not have a hyperlink structure *per se*. It can even be applied to an arbitrary mixture of these content types. Actually, the content of the tagged resources will not have to be accessible in order to manage them in a folksonomy system.

Evaluation We have applied our method to a large-scale dataset from an actual folksonomy system.

Finally, in a preliminary study, we look for similarities between the two approaches (community indentification and ranking/trend detection) in particular comparing FolkRank results with a known cluster detection algorithm (MCL).

Task 3.3 Cluster/community identification

Although research activity on cluster and community detection is planned to begin in the next year, some preliminary results have been obtained after the first year. In particular we have analyzed the structure of the semantic space defined by a given set of resources and discovered the existence of well defined communities of resources corresponding to semantically separated tag clouds.

The study shows how to tame the complexity of the system in order to make manifest the semantic categorization emerging from the anarchic activity of users. To this aim, we use tools and methods developed for cluster detection in complex networks, applied on a specific co-occurrence network builded from the folksonomy data.

Our method try first to reduce frequency effects typical of complex network with the use of similarity metrics, which provide a natural weighting of the network. Then we apply spectral method to identify semantically similar clusters of resources.

We have considered a dataset composed by two different sets, each composed by 200 resources (this number can be easily increased). The two sets are chosen in such a way that the first one only includes resources for which all posts include the tag `design` while the second one only includes resources for which all posts include the tag `politics`. The idea was to artificially construct a dataset with at least two well separated semantic regions. For each resource in the dataset the whole stream of posts is available, i.e. for each resources we have the whole temporal stream of users-tags associations $[U, \{tags\}]$.

The next step has been that of constructing a similarity matrix among resources. The similarity metrics should depend on the frequency occurrence of tags: its specific form has been chosen in order to avoid trivial frequency effects that overestimate the effect of the globally most frequents co-occurrent tags, we explicitly normalize the relative frequencies of tags in by the global frequency. Moreover, when a tag is shared by two resources we try to keep the frequency contribution small.

To this end we define the level of similarity between two generic resources R_1 and R_2 using a TF/IDF weighting procedure (Salton and McGill, 1983). The TF/IDF weight (Term Frequency Inverse Document Frequency) is a often used in information retrieval and text mining and it represents a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In our implementation the tags are playing the role of words. Given two resources, one defines with T_1 and T_2 the set of tags associated with R_1 and R_2, respectively, the Union set as $U := T_1 \cup T_2$, and the Intersection set as $K := T_1 \cap T_2$.

Figure 2.11: Sets of tags considered in evaluating the similarity matrix: T_1 (blue) and T_2 (yellow) are the set of tags associated with resources 1 and 2, respectively. $K := T_1 \cap T_2$ is the set of tags shared by the two resources.

The resulting similarity metrics is shown in Fig. 2.12. Figure 2.12 reports the full matrix of strengths between pairs of resources $w\prime_{R_1,R_2}$ for the full set of 400 resources. The resources are randomly ordered (left) and no structures are evident in this representation.

The problem we have to tackle now is that of finding the sequence of row and column permutations of the matrix of strengths, that permits to visually spot the presence of semantic communities of resources. The goal will be to obtain a matrix with a clear visible block structure on its main diagonal.

To this aim, we made use of spectral techniques recently developed in the context of complex networks (Capocci and Colaiori, 2005; Newman, 2006). The result is shown in Fig. 2.12, right.

The method is based on the identification of the non-trivial eigenvectors of a kind of Laplacian matrix derived from the resource similarity matrix. Given such a set of these non trivial eigenvectors, a very simple way to identify the communities consists in plotting on a multidimensional plot their coordinates. Each axis reports the values of the components of the eigenvectors. In particular each point has coordinates equal to the homologous components for the eigenvectors considered. In this kind of plots communities will emerge as well defined cluster of points. The components involved in each clusters identify the elements belonging to a given community.

Figure 2.13 reports a 3-dimensional scatter plot illustrating the structure of the first three eigenvectors. It is evident the existence of at least 5 well defined communities. It is actually barely visible a sixth community corresponding to the sixth non-trivial eigenvalue.

An interesting question is now whether the communities we have found through the diagonalization of the matrix in Fig.2.12 correspond to semantically separated area in the space of resources. In order to check this point we associate to each community its corresponding tag cloud and we plot it by considering only the 30 most frequent tags. Fig. 2.14 reports the six tag clouds (ordered for decreasing number of resources) where the font of each tag is proportional to the logarithm of the tag frequency.

Despite the intrinsic difficulty of identifying the semantic area defined by a given tag cloud, it is possible to associate, at least to the four main largest communities, well defined areas. In particular the first community could be associated to humor in politics, the second one to visual design, the third one to news in politics and the fourth one to web design. It is evident how the a-priori selfish

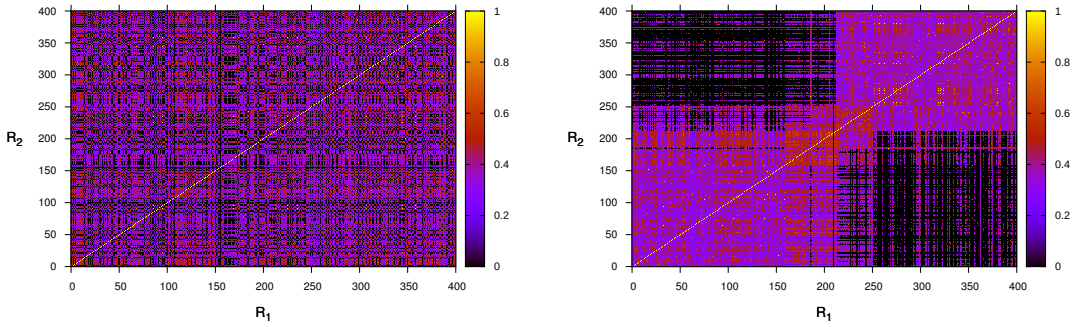


Figure 2.12: Matrix w' of link strengths for the global set of 400 resources. Left: randomly ordered resources. Except for the bright diagonal, whose elements are identically equal to 1 because of the normalization property of the strength w, the matrix appears featureless. Right: indices are ordered by community membership (the sequence of communities along the axes is 2, 4, 6, 5, 3, 1, see Fig. 2.14).

activity of the users is give raise to an effective cooperative behaviors able to structure the semantic space of the resources with semantically meaningful set of tags.

Task 3.4 Semantic inference

With regard to incorporating semantic inference into data analysis, we performed first experiments with a new system which helps users to organize their tags. The resulting Tag-Organizer system (T-ORG) (Abbasi et al., 2007) automatically classifies resources of a tagging system into predefined categories and can help in browsing a particular type of resources available on tagging systems. The resource classification is based on the classification of the tags attached to these resources. If a resource has two tags belonging to two different categories, then the resource is classified as both of these categories.

The core of T-ORG is its classification method T-KNOW (Tag classification using Knowledge On the Web) which is based on the the existing C-PANKOW (Cimiano et al., 2005) system. It provides an unsupervised mechanism for classifying tags in folksonomies. T-KNOW uses Google for finding categories of tags; therefore it does not require any training and can be used for unsupervised classification of tags (like (Cimiano et al., 2005)). It classifies the tags into categories using its pattern library, categories extracted from a given ontology and Google search results. As there might be several results returned by Google against a query posed by T-KNOW, a method is required to select best results on the basis of the similarity between tagging and search results. T-KNOW uses the context of the tag to measure the similarity between Google search results and the tag.

The T-ORG system exploits two different sources of information. On the one hand, it exploits the semantics of the tag co-occurrence which basically correspond to the related tags in the co-occurrence network. The contexts can be used for disambiguating tags by comparison with the categories of other tags in the same context, e.g. the tag “Ford” can stand for the car or the former US president - a vehicle or a person respectively. T-ORG also uses on the other hand semantic background knowledge coming from ontologies found on the web. For example, such an ontology may contain hierarchical knowledge like “Paris” is a city and a city is a location. The semantic background knowledge was used for improving the results. T-ORG was evaluated on a Flickr data set for which encouraging results were achieved.

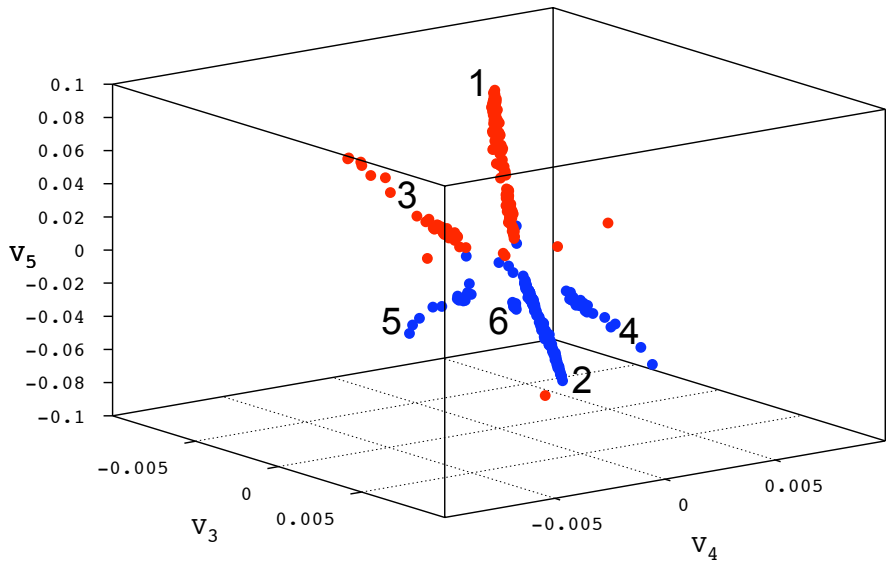


Figure 2.13: Eigenvectors of the matrix Q. The scatter plot displays the component values of the first three non-trivial eigenvectors of the matrix. The scatter plot is parametric in the component index.Five or six clusters are apparent, corresponding to the non-trivial eigenvalues of the matrix. Each cluster, marked with a numeric label,defines a community of "similar" resources (in terms of tags).Blue and red points correspond to resources tagged with *design*and *politics*, respectively. It is important to note that our approach clearly separates the two communities, as well as highlighting a few more finer-grained structures.Tag clouds for the identified communities are shown in Fig. 2.14.

Task 3.5 Cross-Folksonomy Networks

This work will mainly be carried out in the first half of the second year of the project, once data gathering has matured and represented in some accessible and queryable formats. Data from IMDB and Netflix have been already integrated slightly via the movie titles, as described in section 2.1.3. This integration was important for the experiments discussed in section 2.4.2 with respect to predicting movie ratings in Netflix.

Task 3.6 Collaborative tagging and emergent semantics

This task was not envisioned in the original plan of the project, and has been created for the reasons explained in the deviations section.

Task 3.6a Improving Navigation for Images

Although tagging is a simple and intuitive way to organize resources, it also has its limitations. First, people make mistakes while tagging, such as spelling mistakes, or accidental tagging with the wrong tag. Second, there is no solution to cope with homonymy, i.e. to distinguish different meanings of a word. Third, synonymy or different languages can only be handled by tagging data explicitly with all terms.

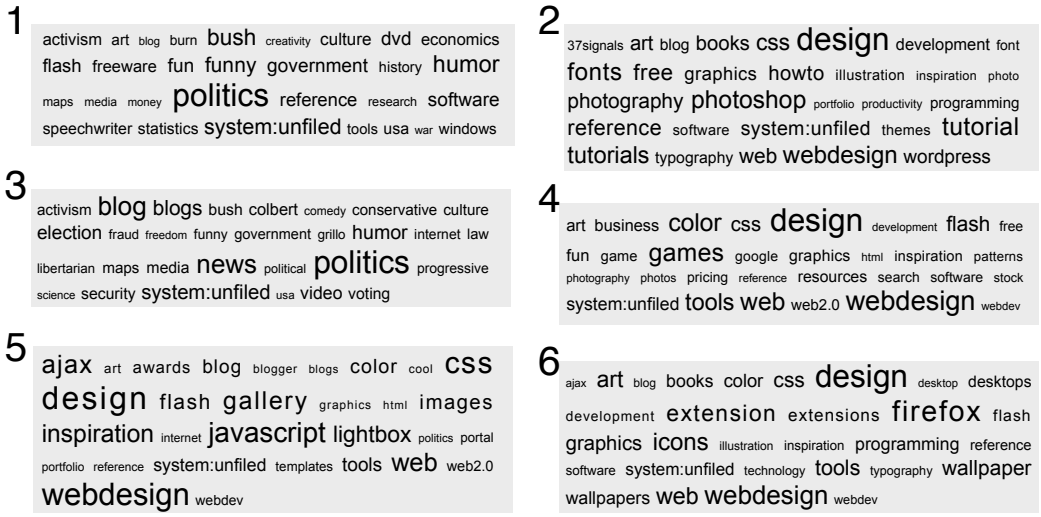


Figure 2.14: Tag clouds for the 6 resource communities identified by our analysis(see Fig. 2.13), ordered by decreasing community size. Each tag cloud shows the 30 topmost frequent tags associated with resources belonging to a given community. Within tag clouds, as usual, the size of text labels increases with the logarithm of the frequency of the corresponding tag. The first two communities (the largest ones) correspond to the main division between resources tagged with *politics* and *design*, respectively. Notice how each tag clouds is strongly characterized by only one of the above two tags. In addition to discriminating the above two main communities, our approach also identifies additional and unexpected communities. On inspecting the corresponding tag clouds, one can recognize a rather well-defined semantic connotation pertaining to each community, as discussed in the main text.

One of the milestones we have reached was to show that content-based analysis can help to overcome some of these limitations. Consider the example in Figure 2.15. On the left are shown a set of images that were tagged *apple* by Flickr.com users. We have a clear case of homonymy: "apple" refers to both the fruit and to the well-known computer company. Through the use of a classifier, based on the simple visual features discussed below, it can be shown that the fruit pictures can be distinguished from the computer images.

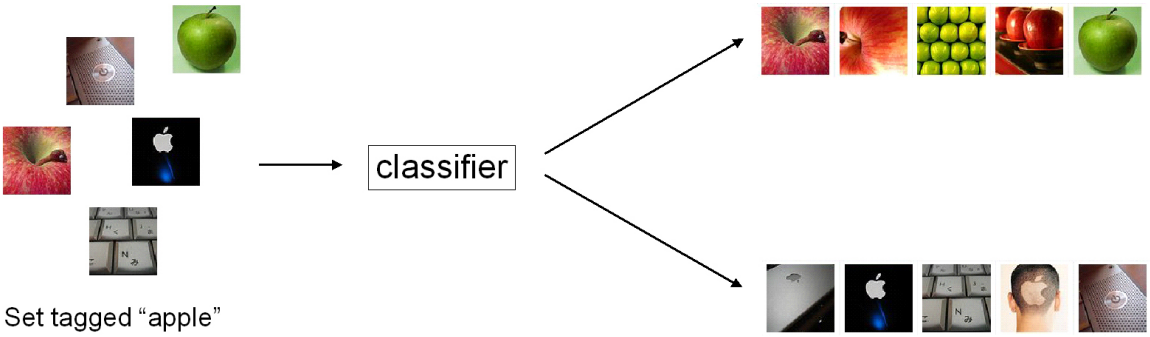


Figure 2.15: A classifier is used to distinguish between two homonyms.

The second example shows the problem of synonymy (Figure 2.16). This particular case shows two tags with the same meaning but expressed in different languages. Starting from an image tagged "bw" (short for black-and-white), a visual similarity search returns several images tagged "noireetblanc" (meaning black-and-white in French). Although no direct relation exists between these two sets of images (nor their tags), the context-based search can introduce the missing

links.

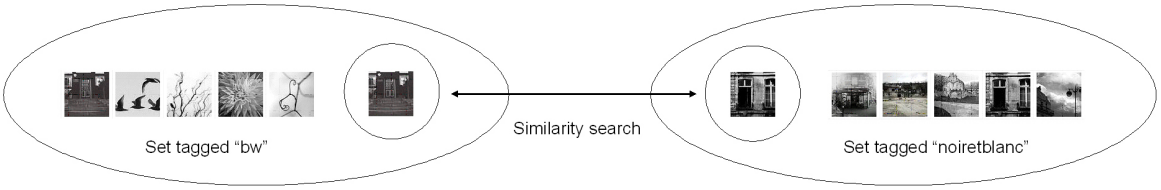
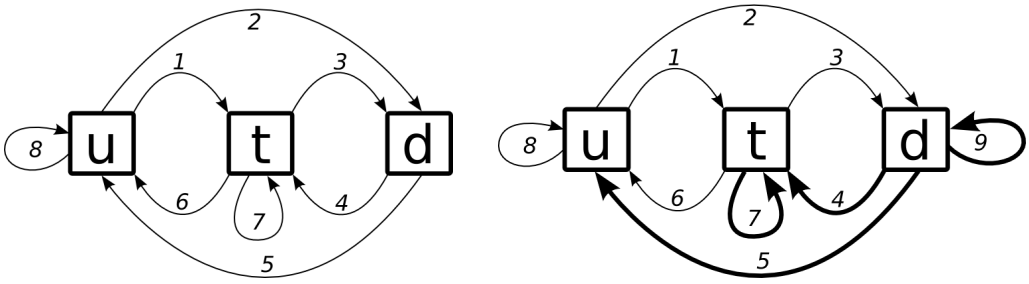


Figure 2.16: Visual similarity is used to link two tags in different languages.

Conceptually, both the similarity search and the content-based classification operate by modifying the links between the nodes in the three-partite network of users, tags, and data (aka resources). Figure 2.17.a shows the navigation possibilities between users (u), tags (t), and data (d) for standard tagging sites. Most of the links are a direct mapping of the relations defined by the tag assignments. The figure shows two additional links: link 7 was introduced using co-occurrence of tags and link 8 is derived from the user's contact list. Both co-occurrence and contact lists are standard features on collaborative tagging sites. Using similarity-based search it is possible to introduce new links (see Figure 2.17.b). In particular, new links between data elements can be established. Data elements that were not tagged or that were very distant in the semantic network of tags, can be brought closer. Taking the analysis one step further, we can also strengthen existing links, or propose new links, between data elements and tags or users. The same content-based approach can also remove links as was shown in the first example above. With the help of a classifier, the user can hide links when the data element if the data elements do not satisfy the visual criteria.



(a) Tagging only (b) Tagging and Similarity Search

Figure 2.17: The navigation map with links between users (u), tags (t), and data (d).

Technical Details To enable the visual similarity search, two features vectors are used: one for color and one for texture. For the color feature, we calculate the first two moments (mean and standard deviation) in RGB color space. In addition, we use the standard deviation between the means of the three color channels. Intuitively, this yields a measure for the colorfulness of an image. The feature has a value of zero for grey-scale images and increases for images with stronger colors. We map the values to a logarithmic scale in order to distribute them more equally. In total, the color feature vector has thus seven dimensions.

Texture refers to the properties that represent the surface or structure of an object. We seek to employ texture features that give a rough measure of the structural properties, such as linearity, periodicity, or directivity of an image. In experiments, we found *oriented Gaussian derivatives* (OGD) to be well-suited for our purposes. This feature descriptor uses the steerable property of the OGD to generate rotation invariant feature vectors. It is based on the idea of computing the "energy" of an image as a steerable function. The first order OGD can be seen as a measure of

"edge energy", and the second order OGD as a measure of the "line energy" of an image. The generated feature vector has 24 dimensions.

When the similarity search is performed on both color and texture, the resulting distance is calculated as a linear combination of the distances in the each feature space.

Besides the positive examples, the user can also give negative examples: they indicate the sort of images that are not sought after. The search then becomes a classification: from the examples images a classifier is build and used to filter the photos. We currently use a nearest neighbor classifier.

The presented work has raised a lot of interest in the tagging research community and has led to two publications (Aurnhammer et al., 2006a,b). Experiments with a generalized framework for feature extraction, based on genetic programming techniques, have led to encouraging results (Aurnhammer, 2007). The feature extraction and similarity-based are available in the Ikoru web site.

Although the "bw" example discussed above yields good results, this is mostly due to the fact that the tag can be grounded in the data. This will not be the case, however, for the majority of the tags. In the next section, we discuss a new approach to train classifiers for higher-level descriptions that are difficult to ground. The approach successfully exploits correlations in the tagging data in combination with content-analysis.

Task 3.6b Improving Automatic Classification of Music

The Sony CSL team has obtained access to a large database of music (about 40,000 titles) that was manually categorized (labeled/tagged) by a group of experts. This categorization process is highly controlled (as opposed to "free" tagging) and an ontology was developed to this aim. About 800 boolean descriptors are available to describe each individual title. In average, each title has 40 descriptors set to "true" and the others to "false. Several studies were performed with this database.

Automatic audio analysis: We try to assess the quality of automatic classifiers trained to categorize music titles with this ontology. Classifiers are trained on a part of the database, and tested on the remaining part. For every descriptor one classifier is constructed (800 classifiers in total). This task raises several issues, notably the definition and automatic construction of "good" training and testing database, to avoid overfitting, but also biases. One of the main contributions of this work is the large-scale evaluation of audio classifiers on "precise" metadata. Furthermore, for this task we rely on a new paradigm to construct classifiers, which makes use of many "ad hoc" audio features (Pachet and Roy, 2007).

Investigation of novel hybrid approaches: Although we can improve the performance of automatic audio classifiers, these classifiers still reach a "glass ceiling" performance (about 70% in average, depending on the nature of the descriptor). This work has led to the definition of a new analysis scheme that uses a hybrid method. First, using a traditional signal-based approach, an attempt is made to ground the descriptors. In general, this will work for only a few descriptors that relate to clearly defined audio features. Second, the initial set of grounded descriptors is used to predict a second set of descriptors based on strong correlative relations between them. The process can be iterated. At each iteration, a new classifier is built using the output of the best classifiers of the previous iteration. While further work is needed, the approach seems to outperform signal-only algorithms by 5% precision on average, and sometimes up the 15% for traditionally difficult problems such as cultural and subjective categories. This work has led to a publication submission (Aucouturier et al., 2007). See also (Campana, 2006a,b). Although this research is conducted using a meticulously annotated database, we foresee that its results will be relevant

to datasets that are annotated through collaborative tagging. This evaluation will be addressed in future work.

2.3.3 Milestones

M3.1 Acquisition of software tools for data analysis (month 5)

Several existing tools and libraries were identified which are useful for different aspects of the tagging data analysis. Examples for tools are Pajek for graph visualization or NET for performing network analysis. A more detailed description will be available as an appendix to D3.1. The development of new tools and scripts used in performing the analysis are also reported in D3.1

Considering the tools support for visualizing large graphs we have been exploring the capabilities of visualization software like Pajek or UCINET. They are capable to visualize some thousand nodes, which is technically only limited by the size of the available main memory. Thats makes them suitable for displaying huge networks of tagging data. The usefulness of such large graphs, however, is another question as the computational overhead increases and the readability and expressiveness usually decreases. Pajek can also factorize a large network into smaller networks which enables refined analysis. Different import and export are supported. Additionally, the tool Text2Pajek allows to convert CSV file into Pajek format which comes in handy when data provided by WP1 needs to be imported.

NET is a tool for statistical analysis of complex networks and was developed in Rome. I takes complex networks as input, usually in plain text files, and computes statistics like degree distribution, betweenness and clustering coefficient. It is also capable of displaying graphs by means of the Graphviz and/or Grip tool.

We also have been investigating the combination or user-tag-resource relations to explore semantic relations between sparsely connected users, tags, or resources that only become visible with the accumulated information from “chaining” such relations. If using a matrix representation for the relations between two of the folksonomy dimensions we can perform the required matrix multiplications with Octave which allows for easy definition of matrices and operation thereon.

T-ORG is a software developed in Koblenz which automatically organize tags into categories. It uses the Google-API and C-PANKOW to retrieve the most frequent associated concepts for tags from the web and then computes their similarity with the tags co-occurrence relation. The organization of the tags into hierarchical categories is done with the help of background knowledge extracted from ontologies.

M3.2 Identification of the key emergent features and global observables relevant for modeling (month 5)

This milestone has been an important check point for both the activities concerning WP3 and WP4. In WP3 we have adapted well-known quantities in network theory to the tripartite structure of folksonomies and identified new important observables to look at. In WP4 we identified important key ingredients to consider in the modeling process. The advances both in WP3 and WP4 are described in the corresponding section.

2.3.4 Deviations and Corrective Actions

SONY-CSL Sony CSL had planned 32 man-months in WP4, on the simulation and control of music and image sharing systems. An extensive body of work has been produced that has been moved to WP3. Indeed, the essence of the work deals more with the analysis of tagging systems than with modeling. In particular, we have studied the relation between tags and content-based

analysis. Our approach starts from a network representation of the tagging data in which edges are introduced, removed, weakened, or strengthened based on the results that stem from an analysis of the content. This line of work thus fits in more naturally in WP3 than in WP4, and a new task, task 3.6, was created in WP3 to enclose it. We have given this work priority because of the potential applications. Progress in this field opens up a plethora of possibilities, from tag suggestions to semantic interoperability (Steels and Hanappe, 2006).

UNIK Because of later hiring, UNIK is slightly behind schedule. With a second full time researcher hired from May 1st, 2007 on, we expect to catch this up soon.

2.3.5 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
3.1	Tools and report for extracting emergent metadata statistics and network metrics. Based on data formats described in WP1, in these tools will provide basic statistical (cf.Task 3.1) and network analysis data (cf. Task 3.2) represented in an agreed and reusable format.	3	31 May 2007	31 May 2007	2	2	**UNI KO-LD**

Mil. No.	Milestone name	WP No.	Date due	Actual/Forecast delivery date	Lead contractor
M3.1	Acquisition of software tools for data analysis.	3	31 Oct. 2006	31 Oct. 2006	**UNI KO-LD**
M3.2	Identification of the key emergent features and global observables relevant for modeling.	3	31 Oct. 2006	31 Oct. 2006	**PHYS-SAPIENZA**

2.4 Workpackage 4 (WP4) - Modeling and simulations

2.4.1 Objectives

The goal of this workpackage is to construct, implement and study specific modeling schemes aiming at reproducing, predicting and controlling the emergent properties seen in the semiotic dynamics orchestrated in on-line communities. We planned in particular a modeling activity at different scales. On the one hand it will be important to construct microscopic models of communicating agents performing language games without any central control. At a different scale we shall

consider more coarse-grained probabilistic models. Several models will be proposed to address specific aspects/scales of folksonomy. The models will allow computer simulation aimed at measuring emergent features to be compared with the results of WP3. The simulations should give an insight in how users select tags, what kind of categories and category structures underlying the evolving system of tags, how categories and tags are related to the objects being tagged, etc. It will also give us information on what kind of more global structures (such as the most frequent tags) can be provided to users to optimize their on-line community infrastructure. The models will require components for assigning or adopting tags, categorizing data, and collective dynamics. However the approach will be to keep the models as simple as possible, identifying the minimal ingredients responsible for the emergent properties. The minimal character of the models should make a more analytical mathematical study feasible. Finally, the output of this WP has the potential to feed back into WP2, specifically to the live social tagging systems developed as part of, in order to experimentally verify the devised control strategies and demonstrate the technological advantage achieved by the present project.

Since the objectives of WP4 are related to developing realistic models for the systems studied in WP1 and WP3, the WP should mainly interest the second and third year of the project. However, a modeling and analyzing activity on the gathered data has begun from the very first months of the project.

Task 4.1 Modeling

This task focuses on developing mathematical models and abstractions of real systems. The basic modelling strategies are the standard ones adopted in Complex Systems Science, namely identifying the minimal required set of basic ingredients that are able to reproduce some selected emergent features of the observed data. Such an approach has the huge benefit of defining an unambiguous measure of success for this task: the quantitative, controllable agreement between the output of the models (produced either by simulation or by analytical approaches) and data from experimental or real systems.

The systems that we plan to study and model exhibit emergent behaviors that arise from the interaction of a large number of elementary entities. In a folksonomy, for example, the shared categorization emerges spontaneously from the distributed, asynchronous - and in principle uncoordinated - tagging activity of web users. That is, there exist at least two scales: a high-level scale where "order" and the emergent dynamics are visible, and a low-level ("microscopic") scale dominated by the individual, uncoordinated activity of agents. At this scale, the activity of agents is seemingly random, and lends itself to be modelled by using probabilistic models. This is similar to what happens in statistical physics, where the appearance of macroscopic order and regularities is explained in terms of the noisy atomic behavior. Because of this, it is natural to use modeling approaches from statistical physics, based on probabilistic and stochastic models. In other areas involving complexity in IT systems, like the study of emergent properties in large-scale technological networks, the use of stochastic models has been greatly successful and we expect them to be similarly useful in the specific context. of this investigation.

Another important objective for the first year corresponds to the deliverable 4.1, which contains a review of theoretical tools and models that we used or plan to use for the study of collaborative social tagging systems.

Task 4.2 Control

Task 4.2.1: Simulation and control on music and image sharing system

Sony-CSL's contribution has been moved from WP4 to WP3. A motivation for this change can be found in the section “Deviations and Corrective Actions” at the end of WP3.

Task 4.2.2: Ontology Learning

This subtask is supposed to start later on. Nevertheless some preliminary studies have been performed by UNI KO-LD.

Task 4.2.3: Simulation and control on bibliographic reference sharing system

This subtask is supposed to start later on.

Task 4.2.4: Recommendations based on Network Analysis

Several e-Commerce systems provide some sort of recommendations to their users. These recommendations are often based on explicit and direct interactions the users make with the systems. Amazon.com for example uses a collaborative filtering technique for its recommendations, which is simply based on tracking what combinations of items users buy. In Last.fm, recommendations are based on user profiles. These profiles are explicitly set by the users, or inferred through their choices of tags. The objective of this task is to experiment with semantic recommendations using the rich and complex networks that are emerging from WP3.

Semantic-based recommendations will be developed that will consider the type of network being analyzed, the type of links incorporated in the network, the semantics of objects within the network, and any temporal information that might be associated with the data.

It is our belief that by harnessing the new levels of data interoperability offered by Semantic Web technology, we will be able to gather information from multiple sources and build large knowledge bases about particular areas of interest (such as music and movies). Traditional resources, such as databases, will be collected and used in conjunction with information extracted from folksonomies to investigate how the two can coexist and be used to better understand both the resources, and how users perceive them.

2.4.2 Progress

Task 4.1 Modelling

In deliverable 4.1 we provided a short review of the theoretical tools for modeling and analysing Collaborative Social Tagging Systems. Particular attention has been devoted to models and tools suited to the analysis of streams of tags. We reviewed some theoretical activity in the field of computational linguistics, whose tools and models could be useful in the study of tag streams in folksonomies.

Task 4.1.1: Stochastic models

On studying *del.icio.us* we started our modeling effort by adopting a tag-centric view of the system, that is we investigated the evolving relationship between a given tag and the set of tags that co-occur with it. In line with our focus on semiotic dynamics, we factored out the detailed identity of the users involved in the process, and only dealt with streams of tagging events and their statistical properties. Specifically, we selected a semantic context by extracting the resources associated with a given tag X and studied the statistical distribution of tags co-occurring with X. Fig. 2.18 graphically illustrates the associations between tags and posts, and Fig. 2.19 reports the frequency-rank distributions for the tags co-occurring with a few selected ones.

The high-rank tail of the experimental curves displays a power-law behavior, corresponding to a generalized Zipf's law (Zipf, 1949) with an exponent between 1 and 2. Since power laws are the standard signature of self-organization and of human activity (Barabasi, 2005; Newman, 2005; Vazquez et al., 2006; Vazquez, 2005), the presence of a power-law tail is not surprising. The observed value of the exponent, however, deserves further investigation because the mechanisms usually invoked to explain Zipf's law and its generalizations (Ferrer i Cancho and Servedio, 2005) don't look very realistic for the case at hand, and a mechanism grounded on experimental data

should be sought. Moreover, the low-rank part of the frequency-rank curves exhibits a flattening typically not observed in systems strictly obeying Zipf's law. Several aspects of the underlying complex dynamics may be responsible for this feature: on the one hand this behavior points to the existence of semantically equivalent and possibly competing high-frequency tags (e.g. *blog* and *blogs*). More importantly, this leveling off may be ascribed to an underlying hierarchical organization of tags co-occurring with the one we single out.

In order to model the observed frequency-rank behavior for the full range of rank values, we introduce a new version of the "rich-get-richer" Yule-Simon's stochastic model (Simon, 1955; Yule, 1925) by enhancing it with a fat-tailed memory kernel. The original model can be described as the construction of a text from scratch. At each discrete time step one word is appended to the text: with probability p the appended word is a new word, never occurred before, while with probability $1-p$ one word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process yields frequency-rank distributions that display a power-law tail with exponent $\alpha = 1 - p$, lower than the exponents we observe in actual data. This happens because the Yule-Simon process has no notion of "aging", i.e. all positions within the text are regarded as identical.

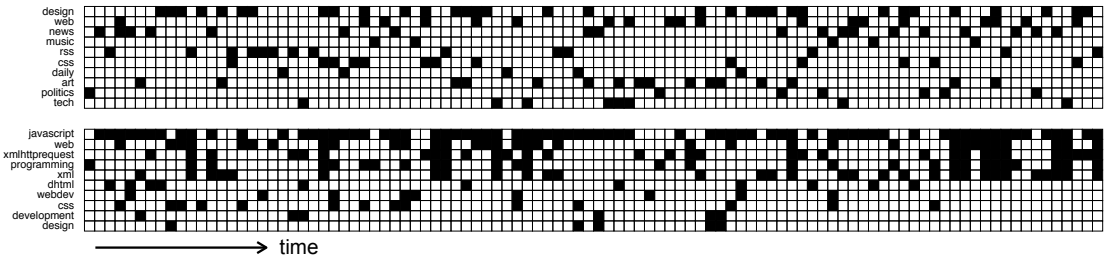


Figure 2.18: Tagging activity: a time-ordered sequence of tagging events is graphically rendered by marking the tags co-occurring with *blog* (top panel) or *ajax* (bottom panel) in an experimental sequence of posts on *del.icio.us*. In each panel, columns represent single tagging events (posts) and rows correspond to the 10 most frequent tags co-occurring with either *blog* (top panel) or *ajax* (bottom panel). 100 tagging events are shown in each panel, temporally ordered from left to right. Only posts involving at least one of the 10 top-ranked tags are shown. For each tagging event (column), a filled cell marks the presence of the tag in the corresponding row, while an empty cell indicates its absence. A qualitative difference between *blog* (top panel) and *ajax* (bottom panel) is clearly visible, where a higher density at low-rank tags characterizes the semantically narrower *ajax* term. This corresponds to the steeper low-rank behavior observed in the frequency-rank plot for *ajax* (Fig. 2.19).

In our modeling we moved from the observation that actual users are exposed in principle to all the tags stored in the system (like in the original Yule-Simon model) but the way in which they choose among them, when tagging a new resource, is far from being uniform in time (see also (Dorogovtsev and Mendes, 2000; Zanette and Montemurro, 2005)). It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones, according to a skewed memory kernel.

We tested this hypothesis with real data extracted from *del.icio.us*. Fig. 2.20 shows the temporal auto-correlation function for the time-ordered stream of tags co-occurring with *blog*, computing over different windows within the experimental stream. The auto-correlation function can be fitted by a power-law of the form $C(\Delta t, t_w) = a(t_w)/(\Delta t + \delta(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor, $\delta(t_w)$ is a phenomenological time scale, slowly increasing with the"age" t_w of the system, and $c(t_w)$ is a base level of correlation due to the finite number of words in the stream. Our modeling strategy consisted of enhancing a simple model for stream generation, embedding

in it a hyperbolic memory kernel of the form stated above.

The stochastic process we introduced (Cattuto et al., 2007) is meant to describe the behavior of an "effective" average user in the context identified by a specific tag, and can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of n_0 words. At a generic (discrete) time step t, a brand new word may be invented with probability p and appended to the text, while with probability $1 - p$ one word is copied from the existing text, going back in time by x steps with a probability $Q_t(x)$ that decays as a power law, $Q_t(x) = a(t)/(x + \tau)$. $a(t)$ is a normalization factor and τ is a characteristic time scale over which recently added words have comparable probabilities. Fitting the parameters of the model, in order to match its predictions (obtained by computer simulation) against the experimental data, we obtain an excellent agreement for all the frequency-rank curves we measured, as shown in Fig. 2.19. This is a clear indication that the tagging behavior embodied in our simple model captures some key features of the tagging activity. The parameter τ controls the number of top-ranked tags which are allowed to co-occur with comparable frequencies, so that it can be interpreted as a sort of "semantic breadth" of a tag.

The above results were reported in Ref. (Cattuto et al., 2007), and an analytical solution of the model for the special case $\tau = 0$ is provided in Ref. (Cattuto et al., 2006).

Task 4.2 Control

Task 4.2.2: Ontology Learning

This subtask was supposed to start later on. Nevertheless two experiments were performed, by UNI KO-LD, where the emergence of patterns in the tagging behavior of users were analyzed. In the first experiment, it was analyzed whether a specific kind of handling compound words was preferred by the users and in the second experiment whether the singular form of a noun is preferred over the plural form. Both experiments were carried out on the Delicious data set collected by the TAGora consortium. For the experiments, UNI KO-LD used the noun categories provided by Wordnet 3.0. During the experiments the vocabulary richness has been analyzed, i.e. the number of distinct tags belonging to one of the Wordnet categories, and the vocabulary usage, i.e. how often was a tag from one of the categories assigned to a resource. Two different patterns have been identified in the relation between the size of a vocabulary and the number of corresponding tag assignments. This relation can be used as an indicator for an increased importance of a topic or rather the existence of a community of users. In the future, UNI KO-LD plans to explore how far the analysis of vocabulary size and usage are suitable for detecting user communities in tagging systems.

Task 4.2.3: Simulation and control on bibliographic reference sharing system

This subtask is supposed to start later on. Nevertheless measures and analysis performed on the network structure of folksonomies (see description of progresses in WP3) suggests the possibility of future control strategies for spam detection and improving navigation. In particular we introduced a network of tag co-occurrence and investigated some of its statistical properties, focusing on correlations in node connectivity and pointing out features that reflect emergent semantics within the folksonomy. We obtained preliminary evidences that simple statistical indicators unambiguously spot non-social behavior as spam.

Task 4.2.4: Recommendations based on Network Analysis

A range of recommendation analysis techniques have already been discussed and studied in the literature (see deliverable D4.4). Some of those techniques have found their way to the industry, where they are being used extensively to provide recommendations to customers. For example,

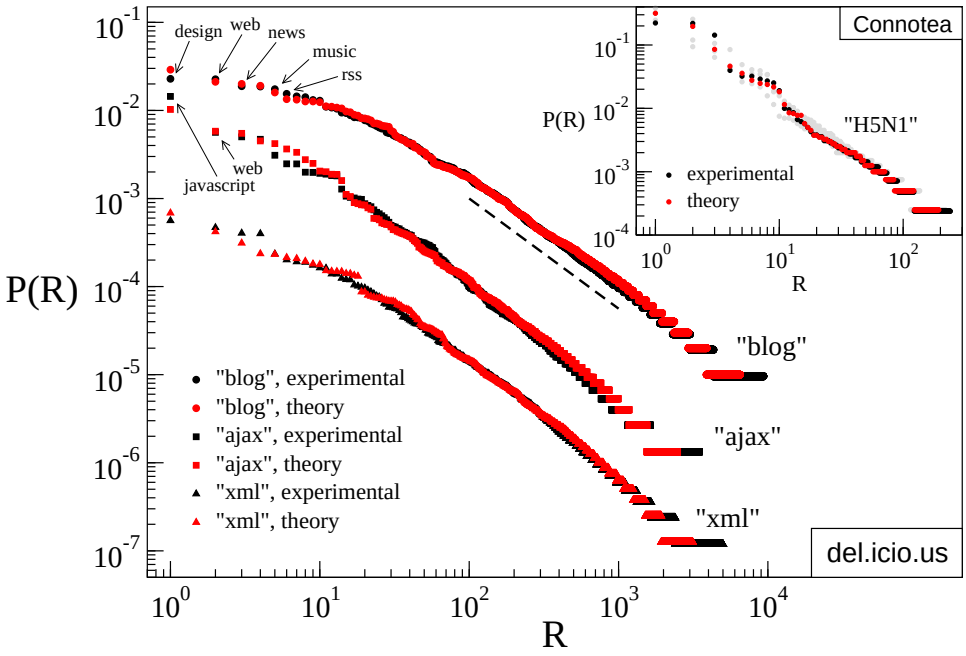


Figure 2.19: Frequency-rank plots for tags co-occurring with a selected tag: experimental data (black symbols) are shown for *del.icio.us* (circles for tags co-occurring with the popular tag *blog*, squares for *ajax* and triangles for *xml*) and *Connotea* (inset, black circles for the *H5N1* tag). For the sake of clarity, the curves for *ajax* and *xml* are shifted down by one and two decades, respectively. All curves exhibit a power-law decay for high ranks (a dashed line corresponding to the power law $R^{-5/4}$ is provided as an aid for eye) and a shallower behavior for low ranks. To make contact with Fig. 2.18, some of the highest-frequency tags co-occurring with *blog* and *ajax* are explicitly indicated with arrows. Red symbols are theoretical data obtained by computer simulation of the stochastic process described in the text (Fig. 2.21). The parameters of the model, i.e. the probability p, the memory parameter τ and the initial number of words n_0 were adjusted to match the experimental data, giving approximately $p = 0.06$, $\tau = 100$ and $n_0 = 100$ for *blog*, $p = 0.03$, $\tau = 20$ and $n_0 = 50$ for *ajax*, and $p = 0.034$, $\tau = 40$ and $n_0 = 110$ for *xml*. Inset: *Connotea* is a much younger system than *del.icio.us* and the corresponding dataset is smaller and noisier. Nevertheless, a good match with experimental data can be obtained for $p = 0.05$, $\tau = 120$ and $n_0 = 7$ (red circles), demonstrating that our model also applies to the early stages of development of a folksonomy. Gray circles correspond to different realizations of the simulated dynamics.

Amazon uses some algorithm to make recommendations on books and other items, and Netflix makes movie recommendations to its users.

To better plan our work on semantic recommendation system, it is important for us to be fully aware of existing techniques, how they compare to each other, their strengths and weaknesses, etc. We also needed to analyze the recommendation techniques that are used by some online systems to acquire some knowledge about what information they use to calculate their recommendations, and where this information is gathered from. Deliverable D4.4 provides a review of existing recommendation strategies and systems. Existing recommendation systems can be roughly divided into two categories: collaborative filtering systems (such as last.fm and Amazon.com), that use similarity in user behaviour to identify new resources of interest; and content-based system (such as NewsWeeder), that use features of the resources, such as term frequencies in a document, to understand the types of items users are mostly interested in. More recent research has investigated

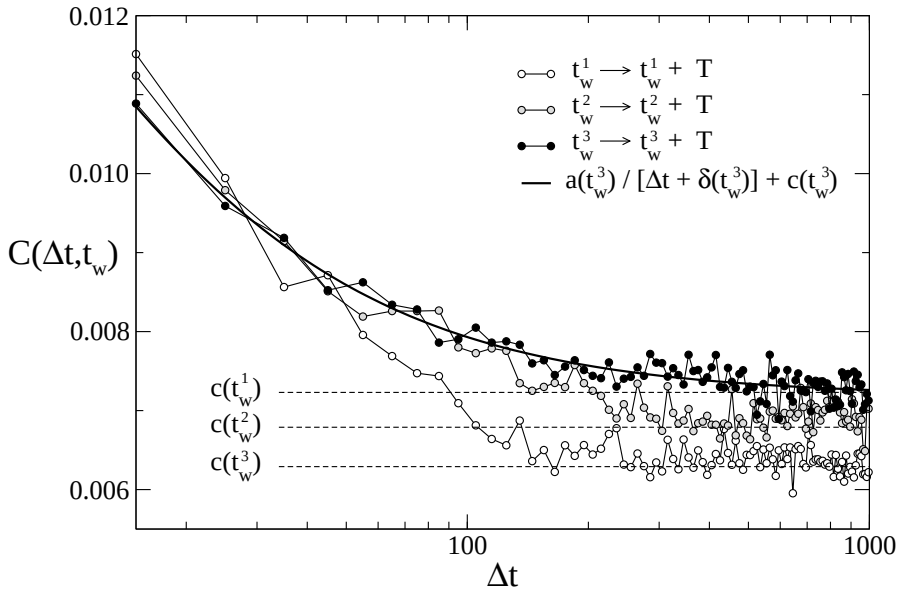


Figure 2.20: Tag-tag correlation functions and non-stationarity. The tag-tag correlation function $C(\Delta t, t_w)$ is computed over three consecutive and equally long ($T = 30000$ tags each) subsets of the *blog* dataset, starting respectively at positions $t_w^1 = 10000$, $t_w^2 = 40000$ and $t_w^3 = 70000$ within the collected sequence. Short-range correlations are clearly visible, slowly decaying towards a long-range plateau value. The non-stationary character of correlations is visible both at short range, where the value of the correlation function decays with t_w, and at long range, where the asymptotic correlation increases with t_w. The long-range correlations (dashed lines) can be estimated as the natural correlation present in a random sequence containing a finite number of tags: on using the appropriate ranked distribution of tag frequencies within each window (see text) the values $c(t_w^1)$, $c(t_w^2)$ and $c(t_w^3)$ can be computed, matching the measured plateau of the correlation functions. The thick line is a fit to the fat-tailed memory kernel described in the text.

a combination of these two strategies through the development of hybrid recommender systems, as well as the gathering of information from other non-traditional sources such as folksonomies.

We have already started experimenting with a recommendation approach that is based on combining information from a traditional movie rental database (Netflix) and a folksonomy about movies (IMDB) (section 2.1.1). We import both data sets into a standard relational database and use string matching to correlate the movie titles in the Netflix data dump with their counterparts in the IMDB data set, providing a way to retrieve IMDB keywords for each Netflix movie title.

To explore the relationship between the way a user rates movies and the keywords that are assigned to movies, we have devised prediction algorithms that guess the rating a user would give to a previously unrated movie based on tag-clouds that depict their interests. For comparison, we also specify a naive average-rating algorithm were the average rating for a movie across all users is used as the predicted rating.

Initial results showed that by building different tag-clouds that express a user's degree of interest, a prediction for a previously unrated movie can be made based on the similarity of its keywords to those of the user's rating tag-clouds. Tests were performed on a subset of the Netflix data comprised of 500 randomly chosen users. For each user, the last 100 movies rated were removed from the training set and used to create a separate test set. Evaluation was performed over the test set by using each of our recommendation strategies to predict the ratings each user would give to the 100 movies in their test set. By comparing the predicted rating to the actual rating given by the

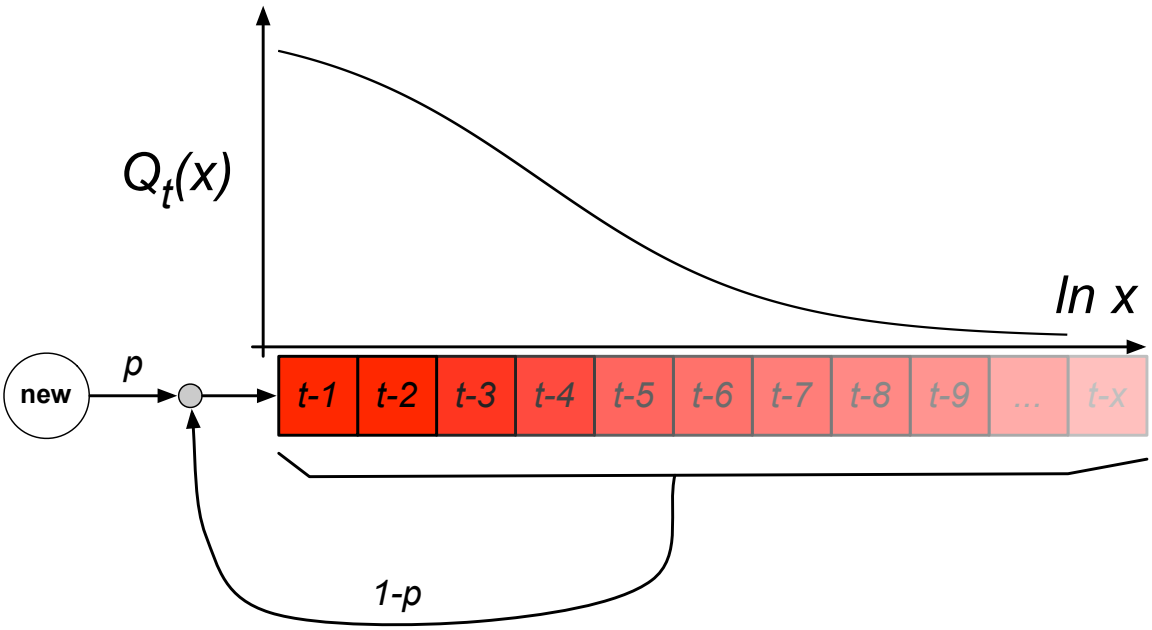


Figure 2.21: A Yule-Simon's process with fat-tailed memory. A synthetic stream of tags is generated by iterating the following step: with probability p a new tag is created and appended to the stream, while with probability $1 - p$ a tag is copied from the past of the stream and appended to it. The probability of selecting a tag located x steps into the past is given by the long-range memory kernel $Q_t(x)$, which provides a fat-tailed access to the past of the stream.

user, we were able to record the root mean squared error (RMSE - a basic measure to express the average distance the predicted rating is from the actual rating) as well as the percentage of correctly predicted ratings. The results of the naive average-rating approach was also recorded to provide a benchmark to compare results to. The naive average-rating approach correctly predicted 36.12% of the movies correctly with an RMSE of 1.131. Our tag-cloud recommendation method correctly predicted 44.15% of the movies with and RMSE of 0.961. More detail of this work can be found in (Szomszor et al., 2007).

2.4.3 Deviations and Corrective Actions

SONY-CSL Sony-CSL's contribution has been moved from WP4 to WP3. A motivation for this change can be found in the section "Deviations and Corrective Actions" at the end of WP3.

UNIK Because of later hiring, UNIK is slightly behind schedule. With a second full time researcher hired from May 1st, 2007 on, we expect to catch this up soon.

2.4.4 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
4.1	(Task 4.1) Review of theoretical tools for modeling and analysing Collaborative Social Tagging Systems.	4	31 May 2007	31 May 2007	2	2	**PHYS-SAPIENZA**
4.4	(Task 4.2) Review of existing recommendation strategies and systems.	4	31 May 2007	31 May 2007	1	1	**UNI-SOTON**

2.5 Workpackage 5 (WP5) - Dissemination and exploitation

2.5.1 Objectives

The objectives for the first year are to communicate the existence and research objectives of the TAGora project to a larger scientific, industrial, and artistic community.

Task 5.1. Project presentation report

This task involves the preparation and publication of a brief project presentation in English accessible to the non-specialist, avoiding technical language, mathematical formulae, and acronyms as much as possible. It is meant to be made available on the World Wide Web through the project web site.

Task 5.2 Dissemination strategies

As tagging becomes an increasingly popular way of classifying large quantities of information in the World Wide Web, it also becomes important to disseminate the results of the research and the products developed by the members of the TAGora project. Since the target audience of this dissemination is not only the scientific community, but also other people interested in tagging, the publication of the results must be done through means that can reach a large audience. Considering this, the main area of dissemination during the first year of the TAGora project was the World Wide Web. By maintaining a web page that includes several dynamic tools such as a blog and, of course, tagging, the TAGora project has aimed to become a point of reference for scientists and other people that want to have a deeper understanding of tagging. However, the dissemination strategies have also included several initiatives outside the Web, such as papers and talks at workshops and conferences, and public demos of the products developed within TAGora.

Task 5.2.1: Explicit dissemination activities on the web
Dissemination on the Web includes the following objectives:

(a) Create a web site in which the public documents generated within TAGora are made available to the public.

(b) Maintain dynamic strategies for informing the Web users about news from TAGora and other relevant items related to tagging.

(c) Provide linkable content (blog posts, event pages) that can be used to establish an effective communication with broader online communities.

Task 5.2.2: The role of applications developed on WP2

The objectives of the applications developed during the first year of the TAGora project, in regard to dissemination, are mainly to reach wide areas of the tagging community. In particular, BibSonomy, developed at the University of Kassel, is already being used by a large number of users. The remaining applications, Ikoru (developed at SONY-CSL) and the peer-to-peer decentralized tagging system developed at the University of Koblenz, are about to be deployed on-line. By providing innovative and easy-to-use approaches to tagging, these applications aim to become popular and well-established tools.

Task 5.2.3: Contribution of SONY-CSL

Every two years, Sony CSL organizes a symposium and open-house, which are a major opportunity to present and demonstrate our work to the scientific community. Two open-house events overlap with the TAGora project.

The Sony laboratory has always sought to interact with the artistic community. These collaborations allow us to explore new interfaces or new usage of collaborative tagging and give us the opportunity to work with small but captivating communities.

Task 5.3 Training activities and outreach

TAGora aims to make public its research results, so that they can be useful in training. Members of TAGora will actively publish and disseminate papers in the scientific community presenting the results of their research. These papers will be listed at the TAGora website. The team will also publish the results in a way that is accessible by non-experts.

Another key dissemination objective of the TAGora project is to reach a wide audience through different information media, such as printed press, radio or television. Members of the project will actively seek to present the TAGora project in different media. In each case, the output is collected and listed at the TAGora website. The applications developed within the project will also be made available to the public in different training and exhibition contexts.

2.5.2 Progress

In this section, the progresses in dissemination achieved by each member of the TAGora project are reported. The reports include progress on dissemination strategies, explicit dissemination on the Web, dissemination of the applications, training activities and outreach.

Task 5.1 (Deliverable 5.1) Project presentation report

The project presentation report, titled “Semiotic Dynamics in Online Social Communities” was prepared with contribution from all Partners of the Consortium. The report is available on the wiki of the TAGora project (`http://wiki.tagora-project.eu/D5.1`), as well as on the project web site for public distribution, at the address `http://www.tagora-project.eu/research/`. The structure of the presentation report is as follows:

- The Vision

- The Challenges
- The Opportunity
- Expected Impact
- Scientific and Technological Objectives
- one section per team, describing the specific expertise, interests and role in the Project of each Partner

Figure 2.22: External face of the TAGora project flyer

In addition to the project presentation report, an additional dissemination item was produced, not initially foreseen in the project dissemination plan. The new item consists of a **project flyer** (see Fig. 2.22 and Fig. 2.23) that was designed to convey basic facts about the TAGora project in a very compact, graphically appealing and non-technical form. 2000 copies of the project flyer were printed and will be distributed at a variety of locations and events. The flyer is available (PDF version) on the private wiki of the TAGora project: `http://wiki.tagora-project.eu/D5.1`.

Task 5.2 Dissemination strategies

The dissemination strategies of the TAGora project have effectively reached an important group of scientists and the general public, thus meeting the objectives stated in the preceding section. The main dissemination initiatives have involved the Web, although many other media were also used. During the project's first year, strategies have included Web-specific dissemination (such as mailing lists and posts), talks at workshops or conferences, the publication of papers and other materials, public tests and exhibitions, and the use of "'traditional"' media (such as newspaper and radio) to communicate activities and research results attained by the TAGora team.

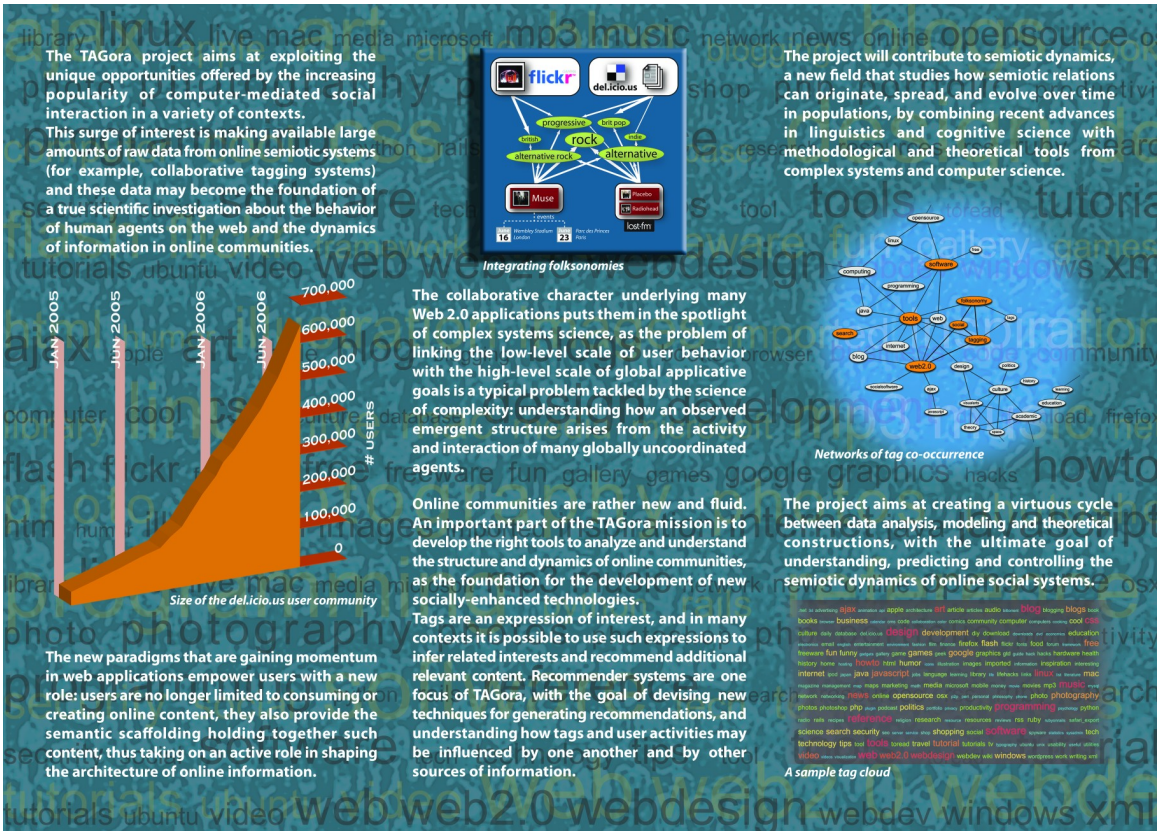


Figure 2.23: Internal face of the TAGora project flyer

The following paragraphs present the progresses made by each member of the project.

Task 5.2.1: Explicit dissemination activities on the web

Although the main way of disseminating the activities and research of TAGora on the Web has been the project's website (please refer to Deliverable 5.2), it has not been the only one. The Intelligence, Agents, Multimedia Group at the University of Southampton has also created a web page in which TAGora is introduced: `http://www.iam.ecs.soton.ac.uk/projects/TAGora.html`. Also, the team members at the University of Koblenz, the Information Systems and Semantic Web group, have developed a webpage that includes research and development done within TAGora. These materials are reachable at the group's webpage: `http://isweb.uni-koblenz.de`.

Task 5.2.2: The role of applications developed on WP2

During the first year of the TAGora project, the applications developed within its context have played a very important role not only as tools for research, but also as dissemination agents. BibSonomy, the social bookmark and publication system developed by the Knowledge and Data Engineering at the University of Kassel, has played a major role in making public some of the research results and technologies developed within TAGora. The project has been announced in a number of mailing lists, and has also been featured in different news sites. For a complete list of the news features in which BibSonomy appears, please refer to `http://www.bibsonomy.org/help/about/press`.

The Information Systems and Semantic Web group at teh University of Koblenz has started to disseminate its peer-to-peer application for decentralized tagging. This application is still receiving some finishing touches before being released to the public as a prototype. A user study will also take place shortly. Although this application has not been released yet, the technologies that are integrated in it have already been the subject of talks in workshops and conferences.

The Ikoru web site recently went online but we have not widely communicated it existence. We hope to gradually attract more users.

Task 5.2.3: Contribution of SONY-CSL

The first combined symposium and open-house was organized in October 2006. To celebrate the laboratory's 10th anniversary a special edition was organised, called "Intensive Science". For this occasion, we organised an exhibition to display some of the artistic projects that members of the lab had realized in collaboration with internationally recognised artists and art schools.

Collaborative tagging found its way into this event through the work of photographer Armin Linke and his students. Seven books were printed that resulted from the tagging experiment at the Venice IUAV university. The exhibition was accompanied with a colourful brochure that contained several view points on the tagging experiment. An additional round table was organized in the exhibition space that gave visitors the opportunity to learn more about semiotic dynamics. More details on this event can be found in Milestone 5.2, below.

The interaction with Armin Linke and his students continue at the Hochschule fur Gestaltung in Karlsruhe, Germany. We participated in several workshops organized by professors Wilfried Kuhn and Doreen Mende and their students of the curator class. The discussions focused on collaborative tagging, semiotic dynamics, collection building, audience participation, and the extension of the museum to the Web.

The Web-based social communication project `http://www.zexe.net`, created by artist Antoni Abad, has also benefited from the research in TAGora. `http://www.zexe.net` involves different misrepresented social collectives broadcasting multimedia contents directly to an unfiltered webpage from mobile phones. The latest version, canal*MOTOBOY, which involves motorcycle messengers in Sao Paulo, Brazil, includes tagging. This refinement to the system was added at SONY-CSL by Eugenio Tisselli as part of the research done in TAGora. The project, canal*MOTOBOY, has been widely announced through different media, and is accessible at `http://www.zexe.net/SAOPAULO`.

Deliverable 5.2 Website for the project

As part of the project presentation task, a web site has been set up as a single reference point for all the public activities of the TAGora project. The second-level domain `tagora-project.eu` was registered and reserved for the TAGora project. A dedicated Linux server was purchased and is currently hosted in the computer room of the PHYS-SAPIENZA team. The server is running the open-source Apache (`http://www.apache.org`) web server and related services. The TAGora web site, `http://www.tagora-project.eu` (Fig. 2.24) is managed by using Wordpress (`http://wordpress.org`), the popular open-source content management system (CMS). The Wordpress installation was highly customized to fit the project needs, both in terms of graphical layout and in terms of content structure. The front page of the web site provides an outline of the TAGora mission and a list of the participating institutes. The main navigation bar (see Fig. 2.24, top) provides access to the following sections:

- **consortium**, `http://www.tagora-project.eu/consortium`: this section gives an overview of the TAGora consortium, highligthing the complementarity of the partners in relation to the project mission. A list of the TAGora teams follows, providing details about the areas of expertise of each team and its position within the international research community.
- **people**, `http://www.tagora-project.eu/people`: this section provides short bios of all the persons involved in the TAGora project.
- **research**, `http://www.tagora-project.eu/research`: gives a more technical overview

Figure 2.24: `http://www.tagora-project.eu`

of the scientific and technological objectives of the TAGora research agenda, focusing on emergent metadata, data analysis, modeling and simulation.

- **products**, `http://www.tagora-project.eu/products`: this section showcases the applications developed in the framework of the project (currently the Ikoru image-tagging application by Sony CSL and the BibSonomy social bookmarking system developed at the University of Kassel).

- **contact**, `http://www.tagora-project.eu/contact`: provides contact information for the project. This includes the email address for project inquiries, `info@tagora-project.eu`, the address of the project coordinator, and information access to the TAGora mailing list for announcements and project news, available at `http://lists.tagora-project.eu/listinfo/news`.

In addition to the above section, three more dynamic website sections are available, whose content is constantly updated to reflect the progress of the project:

- **blog**, `http://www.tagora-project.eu/blog`: this section is a full-blown blog for the TAGora project, co-authored by several project members. The blog is used to announce all kinds of events and news involving the TAGora project, covering scientific results, participation to conferences and workshops, project-sponsored events, press releases, articles in the press and so on. RSS and Atom feeds are provided to allow better exchanges with the blogging community.

- **outreach**, `http://www.tagora-project.eu/outreach`: this section is a subset of the TAGora blog, focusing on blog entries that deal with press releases and articles in the press.

- **publications**, `http://www.tagora-project.eu/publications`: provides a live list of papers published by members of the TAGora project. For each article, the full bibliographic

information is given, together with an PDF version of the paper. The publication list is constantly updated by polling BibSonomy: project members post their new articles to BibSonomy, under the grop "TAGora", and tag them as *tagorapub*. The TAGora website polls the BibSonomy RSS feed for the TAGora group and when it detects new articles tagged with *tagorapub*, it loads them into Wordpress and published them. This kind of integration is an original design deployed by the TAGora project.

All the content included in the above three sections (blog entries, press entries, publication entries) is tagged, and a live tag-cloud is provided in the sidebar of the website (Fig. 2.24, right). Blog posts have editorially given tags, while the tags of publication entries are automatically pulled from BibSonomy. Tags can be effectively used to browse the whole dynamic content of the TAGora website. Standard text search of the website content is also available.

Updated access statistics to the TAGora website are available at `http://www.tagora-project.eu/stats` (password-protected). As of April 2007, the TAGora websites is visited by about 1,100 unique users per month.

Task 5.3 Training activities and outreach

As part of our training activities we organized the following events:

- Gerd Stumme and Christoph Schmitz were co-organizers of a workshop on Semantic Network Analysis that was co-colocated with the European Semantic Web Conference ESWC-2006, June 12, 2006 at Budva, Montenegro.
  (`http://www.kde.cs.uni-kassel.de/ws/sna2006/`)

- Andreas Hotho co-chaired the "Workshop on Web Mining 2006 (WebMine)" that was co-located with the European Conference on Machine Learning/International Conference on Principles of Knowledge Discovery in Databases 2006, September 18th, 2006, Berlin, Germany.
  (`http://www.kde.cs.uni-kassel.de/ws/webmine2006/`)

- Steffen Staab co-organized the 15th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2006, Podebrady, Czech Republic, October 2-6 2006.
  (`http://ekaw.vse.cz/`)

- Vittorio Loreto (PHYS-SAPIENZA) organized the II Bagnovignoni meeting on *Semiotic Dynamics: Grammar*, Bagnovignoni (Siena), Italy, 23-26 October 2006..

- Steffen Staab co-organized the 1st International Conference on Semantic and Digital Media Technologies, Athens, Greece, 6-8 December 2006.
  (`http://www.samt2006.org/`)

- Harith Alani and Gerd Stumme organized a workshop on Social and Collaborative Construction of Structured Knowledge at the World Wide Web Conference 2007.
  (`http://km.aifb.uni-karlsruhe.de/ws/ckc2007`).

- Andreas Hotho co-chaired the Workshop "Bridging the Gap between Semantic Web and Web 2.0" that was co-located with the 4th European Semantic Web Conference ESWC 2007, June 7, 2007, Innsbruck, Austria.
  (`http://www.kde.cs.uni-kassel.de/ws/eswc2007/`)

Additionally we planned the following events:

- Vittorio Loreto (PHYS-SAPIENZA) is the vice-chairman of the XXIII IUPAP International Conference on Statistical Physics, STATPHYS 23, Geona, Italy, 9-13 July 2007. (`http://www.statphys23.org`).

- Vittorio Loreto (PHYS-SAPIENZA) and Luc Steels (SONY-CSL) are organizing the *International School on Complexity: Course on Statistical Physics of Social Dynamics: Opinions, Semiotic Dynamics, and Language*, jointly with Ettore Majorana Foundation and Center For Scientific Culture, Erice, Italy 13-20 July, 2007. (`http://pil.phys.uniroma1.it/erice2007`). In this framework an Atelier devoted to *Modeling Language Evolution with Computational Construction Grammar* has been organized `http://www.csl.sony.fr/erice2007/`.

- Ciro Cattuto (PHYS-SAPIENZA) is organizing a satellite workshop of the European Conference on Complex Systems 2007 (ECCS 2007) on *Social websites: complex dynamics and structure*, Dresden, Germany, 5th October 2007.

Team members of TAGora Harith Alani and Gerd Stumme participated in the organization of a workshop: “Social and Collaborative Construction of Structured Knowledge”, held during the World Wide Web Conference in Banff, 2007. Additionally, other TAGora members have participated in a number of panels and invited talks in different academic contexts.

The training activities held during the first year of the TAGora project have also involved a number of students at the project’s different academic environments. These students have been able to come in contact with state-of-the art research in tagging, and have also provided valuable support for specific tasks. It is the intention of the project to carry on with this activities, and also to produce specific training resources that can be used by professors.

BibSonomy has been presented at several events, including the Intl. Conf. on Data Mining 2006, the Intl. Conf. on Conceptual Structures 2006, at the Research Center L3S, and at the Steering Committee Meeting of the 6FP IP ‘Nepomuk – The Social Semantic Desktop’.

Deliverable 5.3 A *White Paper*

The *White Paper* is mainly intended to describe target problems and grand challenges for Semiotic Dynamics in Online Social Communities, clearly recognized by all, and to openly communicate them to the scientific community.

We are facing a unique opportunity to exploit and give theoretical foundations to the recent, though extremely rapid, developments of emergent semantics in Web-based applications. The joint effort of researchers in many different fields could provide the right trigger to face and tame the challenges of Web Science.

How do microscopic interactions (at the users’ level) affect the macroscopic emergent behaviors of online communities, e.g., how do individual decisions to tag a Flickr photograph, say, affect the information structures that emerge at the macro-scale?

How can we bridge the gap between exploiting natural intelligence (a paradigm commonly referred as Human Computing) and implementing artificial intelligence systems?

In shaping large-scale IT communication systems, how can we bridgethe gap between top-down and bottom up approaches? One of the big failures of user interfaces and human-machine interaction today comes from their lack of adaptivity and the assumption that ontologies and communication conventions can be fixed and imposed from outside.

How will current and emerging resource sharing systems support untrained users in sharing knowledge on the Web in the next years? The knowledge acquisition bottleneck in top-down approaches, i.e., the knowledge transfer from experts to formal systems, should be rephrased here as the *wis-*

dom of crowds issue: is the knowledge aggregation and organization emerging from the uncoordinated activity of millions of users better than a centralized control of few experts?

Trying to answer all these questions is unfeasible at the present. Nevertheless, the *White Paper*, which we developed as Deliverable 5.3, mines the basis for a progressive insight in this new field of Semiotic Dynamics in Online Social Communities.

2.5.3 Milestones

M5.1 Identification of third parties (SMEs) suitable for the deployment of the web-based part of the dissemination plan (month 11).

Members of the TAGora project have independently contacted third parties, communicating them the project's scope and research topics, as well as giving them demonstrations of the applications.

SONY-CSL has meetings to demonstrate the Ikoru system (see Task 2.2) and discuss collaborative tagging with the following people:

- Steve Amagai (Sony Corporation, Tokyo)
- Gerald Reitmayr (Senior Manager E-Business at Digital Imaging Sony Europe)
- Mr Ozawa (Product Planing Dept. VBD, ITCNC, Sony Corporation)
- Akikazu Takeuchi of So-Net (Sony Communication Network Corporation, Tokyo)
- Masanao Tsutsui (Sony CyberShot, Digital Imaging Business Group, Tokyo)
- Dr Sato (Mobile Product R&D Group, Sony Corporation, Tokyo)
- Dr Toshino (PAO Group, Sony Corporation, Tokyo)

The Intelligence, Agents, Multimedia Group at the University of Southampton has made the following contacts:

- A director of Last.fm, who expressed great interest in TAGora in general, and the tasks of trend detection and understanding in particular. As a result of this contact, Last.fm expressed its willingness to contribute with data to the TAGora project.
- A Technical Computing Architect at Redmond Microsoft, who is exploring various social and Web 2.0 issues.

M5.2 The SONY CSL biannual public symposia (2006) in Paris (month 5)

On October 6 and 7, the Sony Computer Science Laboratory (CSL) in Paris celebrated it's 10th anniversary. Over the span of a decade, Sony CSL established itself as a highly innovative place for fundamental multi- disciplinary research with an impact on Information Technology. To celebrate of these achievements at two events were organised under the title *Intensive Science*.

Some of the major scientific breakthroughs of the laboratory are highlighted in the form of a symposium. A following art-science exhibition presented installations resulting from some of the art/science interactions involving members of the laboratory and well known top artists, in the domains of design, music, visual arts, and theatre.

One of the installations on display was "Linking Linke". Photographer Armin Linke has always been interested in how coherent collections could be made from his vast archive of pictures, including personal ones by the viewers themselves. Linke views his pictures as resources for navigating and browsing, but how do you choose from a collection of thousands of images? An experiment

was therefore set up in which art students of the University of Venice could create their personal collections. To help them reach this goal, Melanie Aurnhammer and Peter Hanappe were brought in, two CSL researchers who are investigating similar dynamical processes on the world wide web, including tagging sites such as Flickr.

The experiment proved to be an interesting dialogue between the technologies of social tagging and emergent semantics on the one hand, and the artist Armin Linke and the highly creative students of the University of Venice on the other. At the exhibition, unique collections of Linke's photographs are presented, selected by the students with the use of the Ikoru system provided by the two CSL members.

A discussion panel was organised during the exhibition to talk about the user experiences and Armin's evaluation of the tagging project. A brochure with a vivid description of the exhibition is available.

The event translates into the following figures: 2500 invitations, 400 attendees, 5000 copies of the brochures.

"Linking Linke", M. Aurnhammer, P. Hanappe, A. Linke, M. Brunello, S. Graziani, and the students of the Faculty of Arts and Design, Università IUAV di Venezia, ClaVES undergraduate degree in Visual Arts and Theatre, at La Maison Rouge, Paris 2006.

"Intensive Science", brochure for the 10th anniversary of the Sony Computer Science Laboratory, Paris 2006.

2.5.4 Deviations and Corrective Actions

There have been no deviation or corrective actions during the first year of the project.

2.5.5 Deliverables and Milestones

For each deliverable and milestone please fill in all the missing information: actual delivery date, person months used.

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
5.1	(Task 5.1) Project presentation report.	5	30 Sep. 2006	30 Sep. 2006	1.5	1.5	**PHYS-SAPIENZA**
5.2	(Task 5.2) Website for the project.	5	30 Sep. 2006	30 Sep. 2006	2	2	**PHYS-SAPIENZA**
5.3	A *White Paper* that will describe target problems and grand challenges for for Semiotic Dynamics Systems, clearly recognized by all and openly communicated to the scientific community.	5	31 May 2007	31 May 2007	2.5	2.5	**PHYS-SAPIENZA (All)**

Mil. No.	Milestone name	WP No.	Date due	Actual/Forecast delivery date	Lead contractor
M5.1	Identification of third parties (SMEs) suitable for the deployment of the web-based part of the dissemination plan.	5	31 May 2007	31 May 2007	**SONY-CSL**
M5.2	The Sony CSL biannual public symposia (2006) in Paris.	5	31 Oct. 2006	6 Oct. 2006	**SONY-CSL**

2.6 Workpackage 6 (WP6) - Management

2.6.1 Objectives

Task 6.1. Management

The goals of this WP are: to co-ordinate the administrative and scientific work of the project; to ensure that the management plan is carried out; to monitor progress of the project and provide means to correct deviations from project goals; to ensure that the interface with the Commission runs smoothly; to continually evaluate the project's progress against project and WP objectives, quickly reporting any problems to management; to provide evaluation reports to the Commission as required.

2.6.2 Progress

Task 6.1 Management

The Project Management was carried out by the project coordinator as well as by the Governing Board formed during the kick-off meeting of the project, and node contractors. The project coordinator, Vittorio Loreto, has been and is responsible for the day-to-day co-ordination of the project and has been the main interface between the project and the European Commission. He allocated the financial contribution received from the Commission to the Contractors according to the "Programme of Activities" and the decisions taken by the Consortium. Moreover, the coordinator: (a) verified that the deadline, structure, and content of the deliverables prepared by the contractors are in line with what indicated in the contract, (b) addressed the Project Deliverables to the Commission, after prior validation by the Executive Committee.

The Governing Board (Vittorio Loreto for "Sapienza" University of Rome team, Luc Steels for Sony CSL team, Steffen Staab for the University of Koblenz-Landau team, Gerd Stumme for the University of Kassel team, Harith Alani for the University of Southampton team) was and is responsible for the political and strategical orientation of the project and for any important decision concerning the proper operation of the Consortium.

Contractors (Vittorio Loreto, PHYS-SAPIENZA; Luc Steels, SONY-CSL; Steffen Staab, UNI KO-LD; Gerd Stumme, UNIK; Harith Alani, UNI-SOTON) were and are responsible for: (a) coordinating the research, training and dissemination activities of their node on the basis of the contract and the decision taken by the Governing Board described above, (b) coordinate the preparation of the deliverables and reports for which are responsible, (c) produce a cost statement and an audit certificate every twelve months.

A detailed description of the more important management actions carried during the first year of the project are reported in section 3 of this document. A detailed description of knowledge management, training, and dissemination activities is reported in the Plan for using and disseminating knowledge (D6.2).

2.6.3 Milestones

M6.1 Set up of the project information infrastructure (WWW pages, mailing list, ftp area etc.) (month 3).

As part of the project presentation task, a web site has been set up as a single reference point for all the public activities of the TAGora project. The second-level domain `tagora-project.eu` was registered and reserved for the TAGora project. A dedicated Linux server was purchased and is currently hosted in the computer room of the PHYS-SAPIENZA team. The server is running the open-source Apache (`http://www.apache.org`) web server and related services.
In addition the following services have been activated: email address for project inquiries, `info@tagora-project.eu`, the address of the project coordinator, and information access to the TAGora mailing list for announcements and project news, available at `http://lists.tagora-project.eu/listinfo/news`. A blog has been set up at `http://www.tagora-project.eu/blog/` and a wiki (internal to the Consortium) has been set up at `http://wiki.tagora-project.eu/`.

M6.2-M6.3 Co-ordination and Management Meetings (month 0, 11).

Three Project meetings have been organized during the first year.

Kick-off meeting Rome, June 29/30 2006;

I TAGora meeting Paris, December 13/14 2006;

II TAGora meeting Koblenz, May 15/16 2007;

2.6.4 Deviations and Corrective Actions

No major deviations have to be reported.

2.6.5 Deliverables and Milestones

Del. No.	Deliverable name	WP No.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead contractor
6.1	Provision of reports as required to the Commission.	6	31 May 2007	11 June	1	1	**PHYS-SAPIENZA**
6.2	Yearly Management Report (month 11).	6	31 May 2007	11 June	2	2	**PHYS-SAPIENZA**

Mil. No.	Milestone name	WP No.	Date due	Actual/Forecast delivery date	Lead contractor
M6.1	Set up of the project information infrastructure (WWW pages, mailing list, ftp area etc.).	6	31 Aug. 2006	31 July 2006	**PHYS-SAPIENZA**
M6.2	Kick-off Meeting (month 0).	6	30 June 2006	30 June 2006	**PHYS-SAPIENZA**
M6.2	I Project Meeting (month 5).	6	30 Nov. 2006	14 Dec. 2006	**PHYS-SAPIENZA**
M6.3	II Project Meeting (month 11).	6	31 May 2007	14 May 2007	**PHYS-SAPIENZA**

Chapter 3

Consortium Management

3.1 Consortium Management

The Project Management was carried out by the project coordinator (Vittorio Loreto - PHYS-SAPIENZA -), and by the Governing Board, formed during the kick-off meeting of the project. To foster collaborations among the partners, assure a proper evaluation of progresses and the identification of problems several Project meeting were organized and more specifically:

- **Kick-off meeting** Rome, June 29/30 2006;
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Kassel, October 3-5 2006, focused on WP3.
- PHYS-SAPIENZA and SONY-CSL organized a bilateral meeting in Paris, October 6th 2006, focused on WP4.
- **I TAGora meeting** Paris, December 13/14 2006;
- PHYS-SAPIENZA and UNIK organized a bilateral meeting in Rome, January 31st - February 2nd 2007, focused on WP3.
- PHYS-SAPIENZA and UNI-SOTON organized a bilateral meeting in Rome, March 19-20 2007, focused on WP4.
- UNIK and UNI KO-LD organized a bilateral meeting in Kassel, April 11-12 2007, focused on WP2.
- **II TAGora meeting** Koblenz, May 15/16 2007;

Frequent contacts among participants were also maintained by e-mail, telephone, occasional visits, short and long term visits.

3.2 Problems, deviations and corrective actions

Overall, man-months have been distributed between WPs according to what has been originally planned. No major deviations have to be reported. In some cases we realized that some research activity originally included in a given WP were more related to other WPs. These cases are reported in the deviation section of the corresponding WP.

PHYS-SAPIENZA PHYS-SAPIENZA team used slightly less man-months than the average number of man-months per year. The saved budget has been used to acquire a server for data analysis. A powerful server to be used for data analysis was purchased, for a total expense of 14322.00 euros. The server is a 1U rack-mounted AMD-based machine with 4 dual-core Opteron processors (total of CPU 8 cores), 32 Gb of main memory and very fast (though small) SCSI hard disks. The server has been configured with the Linux/Debian operating system for use by the PHYS-SAPIENZA team and TAGora partners. The large amount of main memory, which accounts for most of the expense, will allow us to cache in main memory the full shapshot of the del.icious and flickr datasets, and perform efficiently global operations over them. Such an amount of memory is also needed to run community-detection algorithms on reasonable subsets of the tri-partite folksonomy network as well as on the tag co-occurrence network.

SONY-CSL SONY-CSL used less man-months than the average estimate due a pregnancy leave of the main post-doc researcher. After this it took some time to hire another person. Globally SONY-CSL shifted a certain number of man-months from WP4 to WP3, as explained in the report of WP3.

UNI KO-LD UNI KO-LD used globally the planned number of man-months but it shifted a certain number of man-months from WP4 to WP2, as explained in the report of WP2.

UNIK Because of later hiring, UNIK is slightly behind schedule. As a consequence UNIK used less man-months than the average estimate. With a second full time researcher hired from May 1st, 2007 on, we expect to catch this up soon.

UNI-SOTON At the beginning of the project UNI-SOTON faced difficulties recruiting a good quality candidate which delayed the task of data gathering by a few months. As a consequence UNI-SOTON used less man-months than the average estimate. As a corrective action, Harith Alani, one of the local TAGora co-PIs dedicated 50% of his time for this project and built some of the needed tools and repositories to gather and store some of the data we require. UNI-SOTON plans to hire one or two final year students over the summer of 2007 to gather more data from specific resources (e.g. Last.fm).

3.3 Project Timetable and Status

Overall the project is progressing as planned and in some cases some Workpackages are in significant advance compared to the objectives originally stated. In particular there have been significant progresses in WP1 (Emergent metadata), WP2 (Applications), WP3 (Data analysis of emergent properties) and WP4 (Modeling and simulations).

Workpackages - Plan and Status Barchart (First year)

Months	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
WP1 - Emergent Metadata						M1.1 M1.2						D1.1 D1.3						
WP2 - Applications						M2.1						D2.1(Task 2.1a) D2.1(Task 2.1b) D2.2 (Task 2.2a) D2.2 (Task 2.2b)						
WP3 - Data Analysis of emergent properties						M3.1 M3.2						D3.1						
WP4 - Modeling and simulations												D4.1 D4.4						M4-1 M4-2
WP5 - Dissemination and exploitation					D5.1 D5.2	M5.2						D5.3 M5.1						
WP6 - Management	M6.2			M6.1								D6.1 D6.2 M6.3						

Workpackages activities have started and are progressing as planned. Deliverables and Milestones have been achieved and delivered in time.

Chapter 4

Other issues

4.1 Co-operation with other projects of the Complex System Initiative

ECAgents - Embodied and Communicating Agents (IST-1940): as partners both of TAGora and the IP ECAgents, PHYS-SAPIENZA and SONY-CSL organized several intiatives, in particular the II Bagnovignoni meeting on *Semiotic Dynamics: Grammar*, Bagnovignoni, October 23-26th October, 2006 and the International School on Complexity, Course on *Statistical physics of social dynamics: opinions, semiotic dynamics and language*, Erice, Italy 13-19 July 2007 (`http://pil.phys.uniroma1.it/erice2007`).

4.2 Co-operation with other European Initatives

NEPOMUK – The Social Semantic Desktop (FP6-027705): NEPOMUK intends to realize and deploy a comprehensive solution for extending the personal computer into a collaborative environment, which improves the State-of–the- Art in online collaboration and personal data management as well as augments the intellect of people by providing and organizing information created by single or group efforts. The objectives include the development of tools for social relation building and knowledge exchange which support knowledge sharing within social communities and the conception of techniques for distributed search and storage of information with the goal of forming a shared knowledge pool within a particular community.

Community management support will be implemented in the Nepomuk social semantic desktop by means of folksonomies. The Tagora project can draw on the first experiences gained and developed in this project as it also focuses on the analysis of network structures within communities of practice. Nepomuk is also a possible application domain for techniques developed in Tagora.

As member of the Research Center L3S, the Knowledge & Data Engineering Group of the University of Kassel is participating in Nepomuk, and serves thus as a bridge between both projects.

NeOn – Lifecycle Support for Networked Ontologies (IST-2005-027595): The aim of NeOn is to support the whole ontology lifecycle. For this purpose, it develops a service-oriented, open infrastructure and a methodology which covers the overall development life-cycle of ontologies. During the maintenance of ontologies one is also interested in re-using and extracting knowledge from sources like document collections, databases or folksonomies.

The University of Koblenz is participating in NeOn and is fostering a knowledge exchange between both projects on techniques which may be used for learning ontological structures from folksonomies and for evaluating ontology learning techniques.

Bibliography

Rabeeh Abbasi, Steffen Staab, and Philipp Cimiano. Organizing resources on tagging systems using t-org. In *Proceedings of the Workshop "Bridging the Gap between Semantic Web and Web 2.0" at ESWC 2007*, June 2007. URL http://www.uni-koblenz.de/~abbasi/publications/T-ORG.pdf.

Harith Alani, Nicholas Gibbins, Hugh Glaser, Stephen Harris, and Nigel Shadbolt. Monitoring research collaborations using semantic web technologies. In *Proc. 2nd European Semantic Web Conference (ESWC)*, pages 664–678, Crete, Greece, 2005.

J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurivé. Signal + context = better classification, 2007. Submitted to the International Conference on Music Information Retrieval (ISMIR).

M. Aurnhammer. Evolving texture features by genetic programming, 2007. Presented at the 9th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing.

M. Aurnhammer, P. Hanappe, and Steels. L. Integrating collaborative tagging and emergent semantics for image retrieval. In *Proceedings WWW2006, Collaborative Web Tagging Workshop*, May 2006a. http://www.csl.sony.fr/downloads/papers/2006/aurnhammer-06b.pdf.

M. Aurnhammer, P. Hanappe, and L. Steels. Augmenting navigation for collaborative tagging with emergent semantics. In *International Semantic Web Conference*, 2006b. http://www.csl.sony.fr/downloads/papers/2006/aurnhammer-06a.pdf.

A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207, 2005.

B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

F. Campana. Une approche généralisée avec eds. Rapport de stage, Université Paris 6, 2006a.

R. Campana. Organisation et catégorisation de la musique populaire par apprentissage statistique. Rapport de stage, ENS Cachan, 2006b.

Andrea Capocci and Francesca Colaiori. Mixing properties of growing networks and the simpson's paradox, 2005. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0506509.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007. URL http://www.pnas.org/cgi/content/short/104/5/1461.

Ciro Cattuto, Vittorio Loreto, and Vito D.P. Servedio. A yule-simon process with memory. *Europhysics Letters*, 76(2):208–214, 2006.

P. Cimiano, G. Ladwig, and S. Staab. Gimme' the context: Context-driven automatic semantic annotation with C-PANKOW. In *Proceedings of the 14th World Wide Web Conference*, pages 332–341, 2005. URL `http://www.aifb.uni-karlsruhe.de/Publikationen/showPublikation?publ_id=889`.

S. N. Dorogovtsev and J. F. F. Mendes. *Phys. Rev. E*, 62:1842, 2000.

R. Ferrer i Cancho and V.D.P. Servedio. Can simple models explain zipf's law for all exponents? *Glottometrics*, 11:1–8, 2005.

Roy T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000. URL `http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm`.

Thomas Franz, Carsten Saathoff, Olaf Görlitz, Christoph Ringelstein, and Steffen Staab. Sea: A lightweight and extensible semantic exchange architecture. In *Proceedings of the 2nd Workshop on Innovations in Web Infrastructure. 15th International World Wide Web Conference (Edinburgh, Scotland)*, 2006. URL `http://www.uni-koblenz.de/~saathoff/publications/sea.pdf`.

Peter Haase, Bjorn Schnizler, Jeen Broekstra, Marc Ehrig, Frank van Harmelen, Maarten Menken, Peter Mika, Michal Plechawski, Pawel Pyszlak, and Ronny Siebes. Bibster - a semantics-based bibliographic peer-to-peer system. In *Proceedings of the WWW'04 Workshop on Semantics in Peer-to-Peer and Grid Computing*, volume 2, pages 99–103, December 2004.

Steven Harris and Nickolas Gibbins. 3store: Efficient bulk rdf storage. In *Proc. 1st Int. Workshop on Practical and Scalable Semantic Systems (PSSS'03)*, pages 1–20, Sanibel Island, FL, USA, 2003.

Andreas Hotho, Robert Jaeschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, 2006a. to appear.

Andreas Hotho, Robert Jaeschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, Lecture Notes in Computer Science. Springer, 2006b. (to appear).

Andreas Hotho, Robert Jaeschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006c. Springer. ISBN 3-540-49335-2. URL `http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006trend.pdf`.

Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981. ISBN 0-201-03822-6.

G. E. Krasner and S. T. Pope. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. *Journal of Object Oriented Programming*, 1(3):26–49, 1988. ISSN 0896-8438.

Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. URL `http://mallet.cs.umass.edu`. http://mallet.cs.umass.edu, 2002.

M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

F. Pachet and P. Roy. Exploring billions of audio features. In Eurasip, editor, *Proceedings of CBMI 07*, 2007. http://www.csl.sony.fr/downloads/papers/2007/pachet-07a.pdf.

Fuchun Peng and Andrew McCallum. Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL*, pages 329–336, 2004. URL http://dblp.uni-trier.de/db/conf/naacl/naacl2004.html#PengM04.

R. L. Rivest. The MD5 message digest algorithm. RFC 1321, Apr 1992. URL ftp://ftp.rfc-editor.org/in-notes/rfc1321.txt. ftp://ftp.rfc-editor.org/in-notes/rfc1321.txt.

G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

Christoph Schmitz, Miranda Grahl, Andreas Hotho, Gerd Stumme, Ciro Cattuto, Andrea Baldassarri, Vittorio Loreto, and Vito D. P. Servedio. Network properties of folksonomies. In *Proceedings of the Tagging and Metadata for Social Information Organization workshop held in conjunction with WWW2007*, 2007.

Andy Seaborne and Christian Bizer. D2rq treating non-rdf databases as virtual rdf graphs. In *Proc. 3rd International Semantic Web Conference (ISWC2004)*, 2004.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425, 1955.

L. Steels and P. Hanappe. Interoperability through emergent semantics. a semiotic dynamics approach. *Emergent Semantics, Special Issue of the Journal on Data Semantics*, 2006.

Martin Szomszor, Ciro Cattuto, Harith Alani, Kieron O'Hara, Andrea Baldassarri, Vittorio Loreto, and Vito D.P. Servedio. Folksonomies, the semantic web, and movie recommendation. In *Workshop on Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference (ESWC)*, Innsbruck, Austria, 2007.

A. Vazquez, J. Gama Oliveira, Z. Dezso, K. I. Goh, I. Kondor, and A. L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006. URL http://www.citebase.org/abstract?id=oai:arXiv.org:physics/0510117.

Alexei Vazquez. Exact results for the barabasi model of human dynamics. *Physical Review Letters*, 95:248701, 2005. URL http://www.citebase.org/abstract?id=oai:arXiv.org:physics/0506126.

D. J. Watts. *Small-worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, NJ (USA), 1999. URL http://www.amazon.com/Small-Worlds-Duncan-J-Watts/dp/0691005419.

G. Udny Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Royal Society of London Philosophical Transactions Series B*, 213:21–87, 1925.

Damián H Zanette and Marcelo A Montemurro. Dynamics of text generation with realistic zipf's distribution. *Journal of Quantitative Linguistics*, 12(1):29–40, 2005. URL http://it.arxiv.org/abs/cond-mat/0212496.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.